المعهد الملكي للثقافة الأمازيغية
ⵉⵔⵉⵎⵏ ⴰⵖⵔⴰⵏ ⵏ ⵜⵓⵙⵙⵏⵜ ⵜⴰⵎⴰⵣⵉⵖⵜ
INSTITUT ROYAL DE LA CULTURE AMAZIGHE

Le Centre des Etudes Informatiques, des Systèmes d'Information et de Communication

Actes de conférence

8ème Conférence Internationale

ⵜⴰⵛⴰⴼⵜ
ⴰⵜⵜⵔⵉⴱⵓⵏⵉⵜ ⵏ ⵓⵙⵎⵓⵜⵜⴰ ⴰ ⵓⵛⵣⴰⵍⴰⵢ

الأمازيغية
وتكنولوجيا المعلومات والتواصل

Technologies d'Information et de Communication pour l'Amazighe

Coordination
Fadoua ATAA ALLAH
Siham BOULAKNADEL

ⵜⴰⵎⴰⵣⵉⵖⵜ ⴰ ⵜⴰⵜⵣⵔⵉⴼⵉⵜ ⵏ ⵜⵉⵡⵎⵓⵜⵉⵏ ⴰ ⵓⵎⵓⵙ

الأمـازيغية وتكنولوجيـا المعـلـومـات والتـواصـل

Technologies d'Information et de Communication pour l'Amazighe

المعهد الملكي للثقافة الأمازيغية
ⵜⴰⵙⴷⴰⵡⵉⵜ ⴰⵔⵡⴰⵙ ⵏ ⵜⵓⵙⵙⵏⵉ ⵜⴰⵎⴰⵣⵉⵖⵜ
INSTITUT ROYAL DE LA CULTURE AMAZIGHE

Centre des Etudes Informatiques,
des Systèmes d'Information et de Communication

Actes de conférence

8ème Conférence Internationale

ⵜⴰⵎⴰⵣⵉⵖⵜ ⴰ ⵜⴰⵜⵣⵔⵉⴼⵉⵏⵉⵣⵜ ⵏ ⵜⵉⵍⵢⴰⵣⵉⵜ ⴰ ⵜⵎⵣⴰⵡⴰⵏ

الأمـازيغية وتكنولوجيـا المعلـومـات والتـواصـل

Technologies d'Information et de Communication pour l'Amazighe

Coordination
Fadoua Ataa Allah, Siham Boulaknadel

Publications de l'Institut Royal de la Culture Amazighe
Centre des Etudes Informatiques, des Systèmes d'Information et de Communication
Série : Colloques et séminaires N° 52

Titre
Technologies d'Information et de Communication pour l'Amazighe

Coordination
Fadoua Ataa Allah, Siham Boulaknadel

Conception
Nadia Kiddi (Unité de l'édition)

Editeur
Institut Royal de la Culture Amazighe

Imprimerie
Edition et impression Bouregreg - Rabat

Dépôt légal
2015 MO 1163

ISBN
978-9954-28-184-0

ISSN
2421-9711

PREFACE

La sélection d'articles publiés dans ce recueil constitue les actes de la 8ème édition de la conférence Internationale sur les Technologies de l'Information et de Communication pour l'AMazighe (TICAM) qui s'est tenue les 26 et 27 novembre 2018, à l'Institut Royal de la Culture Amazighe. Cette conférence est le rendez-vous privilégié de la communauté marocaine qui s'intéresse aux problématiques liées à l'informatisation des langues. Chercheurs académiques, industriels et étudiants s'y retrouvent pour échanger sur des thématiques de recherche propres à l'enseignement-apprentissage via la technologie et le traitement automatique des langues.

À l'heure du numérique, l'informatisation des langues est devenue un facteur déterminant de son usage et de sa survie. En effet, les langues qui ne disposent pas de logiciels de traitement automatique du langage naturel ne pourront plus aspirer à être des langues véhiculaires dans les domaines des sciences, des techniques et des affaires. Cette dimension des industries de la langue exige en particulier de garantir la disponibilité de ressources linguistiques (terminologies, dictionnaires, grammaires) permettant la réalisation d'applications industrielles.

La conférence a reçu dix-sept soumissions, d'auteurs provenant de 3 pays différents. Dans cette édition, les thèmes abordés par les auteurs couvrent différents aspects qui sont au cœur de la communauté tels que l'enseignement-apprentissage médiatisé par la technologie, le traitement des données textuelles, la reconnaissance optique des caractères et le traitement de la parole. Deux personnalités dont les travaux sont liés à l'informatisation des langues ont accepté de présenter certains aspects de leurs domaines de recherche lors de cette conférence. Rachida Ajhoun, Enseignante-Chercheure à l'ENSIAS-Université Mohamed V, a abordé les enjeux du Digital Learning entre Hier et Aujourd'hui. Guy De Pauw, Chercheur à l'Université d'Antwerp, a exposé sa vision sur l'analyse textuelle des langues peu dotées.

Nous remercions vivement les auteurs pour leurs contributions, les conférenciers invités pour avoir honoré la conférence de leur présence, les modérateurs des sessions pour leur investissement ainsi que le comité scientifique pour la qualité de leurs relectures. Nous remercions chaleureusement le comité d'organisation pour son travail efficace. Enfin, nous remercions l'Institut Royal de la Culture Amazighe pour son soutien à la conférence.

Comité scientifique

Ait Kerroum Mounir (ENCG, Kénitra)
Aoughlis Farida (UMMTO, Tizi Ouzou)
Ataa Allah Fadoua (IRCAM, Rabat)
Bennani Samir (EMI, Rabat)
Bouikhalene belaid (FST, Béni Mellal)
Boulaknadel Siham (IRCAM, Rabat)
Bouyakhf Houssaine (FSR, Rabat)
Cavalli Sforza Violetta (AUI, Ifrane)
El Marraki Mohamed (FSR, Rabat)
El Qadi Abderrahim (EST, Meknes)
Fadouli Nouredine (EMI, Rabat)
Fadili Hammou (CNAM, Paris)
Fakir Mohamed (FST, Beni Mellal)
Idrissi Najlae (FST, Beni Mellal)
Khalidi Idrissi Mohamed (EMI, Rabat)
Mouradi Abdelhak (MESRSFC, Rabat)
Morin Emmanuel (LINA, Nantes)
Pognan Patrice (INALCO, Paris)
Rami Salim (FPS, Safi)
Rosso Paolo (UPV, Valence)
Rguig Mohand (USMBA,Sais Fes)
Semmar Nasredine (CEA, Paris)
Sidir Mohamed (UPJV, Amiens)
Tigziri Noura (UMMTO, Tizi Ouzou)
Yousfi Abdellah (FSJES, Rabat)
Zerouali Mohamed (CFIE, Rabat)
Zock Michael (Univ-Marseille, Marseille)

Comité d'organisation

Centre des Etudes Informatiques des Systèmes d'Information et de Communication

Table des matières

Digital Learning : les enjeux entre hier et aujourd'hui

Abstract

Le monde de l'apprentissage, à l'image de la société et à l'instar du monde de l'information, est en pleine mutation. Bousculé par le digital, les réseaux sociaux et l'évolution des pratiques socio-culturelles tentent de s'adapter. Cette transformation ne se fait pas sans questionnement sur l'efficacité pédagogique des nouveaux dispositifs (Mooc, serious games, mobile learning, classe inversée, ...) et sur les risques potentiels qu'ils induisent sur l'écosystème pédagogique. Quelle est la transformation de l'apprentissage de demain? Le système éducatif, tel que nous le connaissons aujourd'hui, existera-t-il encore dans les années à venir? Les nouvelles technologies telles que la réalité augmentée et la réalité virtuelle, les robots d'apprentissage et les nouveaux appareils mobiles amélioreront-elles les expériences d'apprentissage? A quoi ressembleront le futur enseignant et le futur apprenant?

Biographie

Rachida AJHOUN

est professeur de l'enseignement Supérieur à l'ENSIAS (Ecole NationaleSupérieure d'Informatique et d'Analyse des Systèmes) Université Mohammed V Rabat,Maroc. Elle a obtenu son diplôme de doctorat d'état en sciences informatiques (adaptabilité des cours à distance) de l'Ecole Mohammadia d'Ingénieurs, Maroc en 2001. Elle est membre fondateur du e-Learning Center de l'Université Mohamed-V-Rabat. Elle a été nommée directrice de ce centre de 2011- 2015. Responsable de l'équipe de recherche LeRMA(Learning and Research of Mobile Age) à l'ENSIAS. Elle est responsable de plusieurs projets (recherche et formation) nationaux et internationaux en e-Learning. Pr. AJHOUN a participé à plusieurs projets de formation en e-Learning et MOOCs. Elle est responsable du projet « Production du premier MOOC marocain » soutenu par l'AUF en 2013. Conseillère technique au sein du Ministère de l'enseignement supérieur, de la recherche scientifique et de la formation des cadres et responsable du projet «Ressources Pédagogiques Numériques» du programme e-SUP durant 2013-2014.Elle est membre fondateur de l'association GUIDE (Global Universities in Distance Education) crée en 2005 et représentante de la région Afrique entre 2008-2012.Elle estmembre seniorde l'IEEE, membre du comité consultatif de la conférence EDUCON. Elle est l'auteur et co-auteur de plus de soixante articleset des communicationset auteurd'un livre et de 3 chapitres de livres sur l'e-learning. Directeur de 14 thèses en digital Learning dont 7 soutenues entre 2009 et 2017.

Industrial Text Analytics for resource-scarce languages

Abstract

Social media have allowed people to freely share their views on products, entertainment, politics, current events and… each other. This magnitude of opinions advances at an incredible pace, but managing the knowledge that is contained in this stream of unstructured data is no longer possible through mere human means. w ith automatic text analytics, however, we now have the technology to automatically collect and monitor opinions in order to turn language into societal insights. Unfortunately, the industry typically only focuses on commercially interesting languages, while underrepresented language communities perhaps may benefit the most from such technology. In this talk, we will present some of our ongoing projects and outline our attempts at porting our technology to resource-scarce language groups such as Amazigh

Biographie

Guy DE PAUW

Guy De Pauw has been working as a language engineer for more than 20 years and has extensive experience developing robust text analytics applications for a wide range of languages. As co-founder and CEO of the University of Antwerp spin-off Textgain, he is now bringing artificial intelligence to the market that can extract knowledge from documents in the context of automation, trend analysis and user profiling. At the cross-section between industry and academia, Guy is always searching for industrial applications for the state-of-the-art in cognitive computing.

TICAM'18, p. 11 - 20

MOOC-IRCAM pour l'apprentissage de la langue amazighe - opportunités et défis

Youness CHAABI, Siham BOULAKNADEL, Fadoua ATAA ALLAH
CEISIC, Institut Royal de la Culture Amazighe, Rabat, Maroc
{chaabi,boulaknadel,ataaallah}@ircam.ma

Résumé

Cours en ligne ouvert et massif (CLOM), connu sous sa dénomination anglaise Massive Open Online Courses (MOOC), est un développement assez récent dans le domaine de l'enseignement à distance. L'accès à l'éducation ouverte, au contenu ouvert et aux Ressources Educatives Libres (REL) attire de plus en plus l'attention des apprenants. L'arrivée des MOOC a changé radicalement l'idée de l'éducation et a orienté les apprenants vers des cours éducatifs distribués, participatifs et en même temps soutenant l'apprentissage en réseaux.

Les MOOC de langues (LMOOC) ont été récemment ajoutés à la liste croissante des cours ouverts. Pour les apprenants, les MOOC offrent une alternative innovante et peu coûteuse à l'apprentissage formel et traditionnel. Pour les concepteurs de cours, ce modèle d'apprentissage émerge des questions importantes concernant le montage du nouvel environnement d'apprentissage et l'adoption d'une approche pédagogique particulière pour soutenir l'expérience d'apprentissage.

Les auteurs de cet article offrent un aperçu de leurs propres expériences dans la conception et le montage d'un MOOC pour l'apprentissage de la langue amazighe pour les adultes sur MOOC-IRCAM. Cette étude explore les opportunités et les défis qu'ils ont rencontrés et le lien avec les recherches existantes.

Mots-clés : Cours en ligne ouvert et massif, MOOC, Apprentissage des langues étrangères, Education aux langues, Ressources Educatives Libres, apprentissage des langues en ligne.

1. Introduction

De nombreuses recherches ont montré que le web est un domaine riche en mettant la vraie pratique de la communication, et explore de nouvelles formes pour exercer sa compréhension et sa fluidité. Plus précisément, le web 2.0 est participatif, immédiat, authentique et engage la communauté. Toutes ses caractéristiques sont essentielles dans le processus d'apprentissage des langues, et font du web 2.0 un environnement prometteur pour l'apprentissage des langues.

Ces dernières années, grâce au web et développement des ressources éducatives libres (REL), le nombre de ressources éducatives disponibles gratuitement en ligne a augmenté de façon exponentielle (McGreal, 2013). Les REL proposent diverses caractéristiques telles que

l'accès libre, l'utilisation et la réutilisation des ressources, l'ouverture et tous les principes communs qui pourraient favoriser et améliorer la construction de réseaux d'apprentissage à distance (McGreal, 2013).

Les MOOC jouent un rôle important dans le développement de l'apprentissage à distance, qui a évidemment suivi la montée en puissance de l'éducation en ligne, et le développement des universités en libre accès à travers le monde (Siemens, 2013). Ce nouveau phénomène éducatif a changé radicalement l'idée de l'éducation et a orienté les apprenants vers des cours éducatifs ouverts, distribués, participatifs et en même temps soutenant l'apprentissage en réseaux. On peut dire que le premier MOOC a été créé et présenté par George Siemens et Stephen Downes en 2008 (Parr, 2013) pour tester l'approche connectivisme. Cette approche pédagogique focalise sur les apprenants qui collaborent pour co-construire et distribuer des connaissances à travers des réseaux, en tant que pratiquants dans une communauté. Quatre ans après, selon le New York Times, 2012 a été « l'année du MOOC » (Siemens, 2013 ; Pappano, 2012), tandis que « CourseEra », le plus grand fournisseur de MOOC, a enregistré 2,8 millions d'apprenants en mars 2013, des collaborations avec des universités prestigieuses et des centaines de cours ouverts en plusieurs langues (Bárcena *et al.*, 2014). Récemment, de nombreuses organisations et institutions éducatives privées et publiques dans le monde ont décidé de profiter de l'occasion de faire des investissements dans la conception pédagogique, le développement de cours et de plateformes soutenus par les technologies émergentes.

L'intérêt croissant pour des opportunités d'apprentissage en ligne massif et ouvert ouvre de nouvelles opportunités pour l'apprentissage des langues. D'un côté, les technologies de web 2.0 et 3.0 offrent un environnement d'apprentissage des langues interactif et authentique, qui favorise la collaboration, l'apprentissage autonome et libre (Hurd, 2016), l'évaluation par les pairs, etc. D'un autre côté, il existe un besoin éducatif énorme pour apprendre la langue amazighe, tant au niveau national qu'international, en raison de la mondialisation de la société dans laquelle nous vivons, travaillons et apprenons.

Dans cet article, nous présentons une revue de la littérature sur les MOOC et nous discuterons des caractéristiques de base d'un cours d'apprentissage en ligne réussi et de sa transition vers un cours de langue en ligne ouvert à tous. Nous présenterons les étapes clés pour la conception d'un environnement d'apprentissage efficace pour les MOOC. Ensuite, nous poursuivrons la description de notre expérience dans la conception d'un MOOC pour l'apprentissage de la langue amazighe pour les adultes. Les opportunités et défis de cette expérience seront présentés et discutés. Enfin, nous présentons quelques conclusions finales et des perspectives de cette recherche.

2. Etat de l'art

L'enseignement à distance a traditionnellement donné accès aux programmes d'enseignement pour les apprenants isolés. Le matériel éducatif avait généralement un format prédéfini (des textes, des cours audio et/ou vidéo) pour encourager l'apprentissage. De nos jours, avec l'avènement des technologies de l'information et de la communication, cette perspective a radicalement changé. Même si la distance entre apprenant et tuteur demeure une caractéristique fondamentale de l'éducation à distance, l'interactivité et la collaboration jouent un rôle important dans cette nouvelle réalité éducative en ligne. En fait, nous pouvons

décrire cet environnement éducatif en ligne comme riche, ouvert, participatif, distribué et constamment en faveur de l'idée de l'apprentissage en réseau.

Les principaux avantages des MOOC sont l'autonomie, la diversité, l'ouverture et l'interactivité (Downes, 2012a ; Downes, 2012b). Le socio-constructivisme, l'apprentissage collaboratif et le connectivisme sont les principes théoriques qui sous-tendent le développement des MOOC et leur avenir, dans le cadre desquels les apprenants soutiennent la communauté d'apprentissage par l'interaction sociale et l'engagement actif dans le processus d'apprentissage. Cependant, il existe différentes approches pour la conception et la mise en œuvre de MOOC dérivant des principes théoriques distinctifs. Depuis 2012, selon Stephen et de nombreux chercheurs et spécialistes (Rodriguez, 2013, Rodriguez, 2012, Mackness, 2010), l'interprétation de ce que sont les MOOC n'est pas simple. En fait, il existe deux types de MOOC : Les cMOOCs et les xMOOCs. Généralement, les MOOCs partagent deux caractéristiques clés communes : l'ouverture et l'évolutivité. Cela signifie que n'importe qui peut avoir un accès libre à un cours en ligne gratuitement pour soutenir un nombre indéterminé de participants (Downes, 2011). La principale raison de cette catégorisation était les différents fondements pédagogiques de ces cours.

Les cMOOCs soutiennent les principes de l'autonomie et le connectivisme (La connaissance est distribuée à travers un réseau de connexions), et en conséquence, l'apprentissage consiste en la capacité de construire et de traverser ces réseaux (Downes, 2012a ; Downes, 2012b). Dans les cMOOCs, le contenu des cours n'est pas considéré comme l'objet de l'apprentissage mais plutôt comme un outil d'aide à l'apprentissage et qui active l'engagement des apprenants dans une communauté. cMOOCs maintiennent les connaissances émergents de l'interactivité, la diversité des réseaux sociaux et l'apprentissage par les pairs (Mackness, 2010). Dans ce modèle, qui met l'accent sur l'apprentissage social ouvert et l'indépendance des apprenants, les apprenants collaborent pour réunir et distribuer les connaissances à travers différents réseaux, tandis que les tuteurs démontrent des scénarios pédagogiques, stratégies et techniques « l'approche, le langage et la vision du monde » (Downes, 2011). Dans ce contexte, les cMOOC offrent une plate-forme pour explorer de nouvelles pédagogies au-delà des salles de classe traditionnelles (Yuan et Powell, 2013) et introduisent des activités de collaboration, d'apprentissage et d'alimentation des ressources.

Cependant, les xMOOC sont basés sur la pédagogie cognitive-behavioriste. Ils soutiennent un modèle axé sur la technologie et centré sur le tuteur qui établit une relation tuteur-apprenant et apprenant-apprenant (Rodriguez, 2013). Les apprenants sont encouragés à suivre une série d'activités concrètes et adaptées à leurs besoins qui offrent une rétroaction automatisée. Les xMOOC offrent des cours universitaires de qualité à un public étendu. Ils sont construits dans un contenu bien structuré et suivent une approche instructiviste selon laquelle les cours sont conçus avec des objectifs d'apprentissage et d'enseignement spécifiques intégré dans les ressources du cours en ligne (Ferguson *et al.*, 2016). Dans les xMOOC, l'équipe pédagogique est responsable de la structuration et la prestation des cours ainsi que sa conception. Comme l'indique Littlejohn (Littlejohn, 2013), ces MOOC ne nécessitent pas un bon niveau d'interaction entre les apprenants. La majorité de ces types des MOOCs, plutôt que de suivre un cadre connectiviste, utilise une approche pédagogique plus traditionnelle et instructiviste (Kennedy, 2014 ; Staubitz *et al.*, 2015) qui se concentre sur les théories de l'autonomie et de l'autorégulation des apprenants.

Dans l'enseignement des langues, le modèle éducatif d'un cours de langue en ligne ouvert et massif doit être étudié avec circonspection. Ce nouveau modèle s'appelle LMOOC (Bárcena *et al.*, 2014). La première contribution majeure est l'analyse des problèmes théoriques et méthodologiques liés aux LMOOC.

Les LMOOC est une nouvelle approche pédagogique basée sur une synthèse de cMOOC et xMOOC, exploite le potentiel de l'approche en réseau ainsi que la pédagogie structurée de l'enseignement supérieur. Le 'L' représente la responsabilité individuelle, l'interaction, les relations interpersonnelles, l'innovation et l'implication. Les LMOOCs peuvent être efficacement conçus pour faciliter le développement de compétences linguistiques et communicatives, massives et hautement hétérogènes, dont le principal intérêt est d'apprendre une langue étrangère. Grâce à des outils éducatifs, les apprenants utilisent leur propre environnement d'apprentissage personnel (EAP) pour gérer leur apprentissage et engager une conversation avec d'autres apprenants.

La formation à distance ouverte et massive des langues, soutenue par des technologies émergentes, offre des défis pédagogiques, mais en même temps créent de nombreux problèmes. Les spécialistes soulignent que l'un des principaux défis auxquels sont confrontés les LMOOCs est que l'apprentissage d'une langue est généralement basé sur les compétences plutôt que sur la connaissance, et que la pratique nécessite d'apprendre avec les autres, alors que la majorité des LMOOCs existants suivent une approche instructiviste qui ne favorise pas la collaboration. Par contre, l'objectif est donc de favoriser un environnement qui améliore l'apprentissage social en incluant une gamme d'activités et d'outils qui soutiennent la discussion et la collaboration entre apprenant-apprenant et tuteur-apprenant.

Dans les sections suivantes, les auteurs de cet article essaieront d'expliquer quand un LMOOC peut être un environnement efficace d'apprentissage. Par la suite, ils discuterons les défis et les opportunités qu'ils ont rencontrés dans la conception et la mise en place du MOOC-IRCAM pour l'apprentissage de la langue amazighe pour adulte sur la plate-forme OpenEdx, et comment ils se rapportent à ces différentes approches cMOOCs, et xMOOCs.

3. Étapes clés pour la conception d'un environnement d'apprentissage efficace pour les LMOOC

Dans l'enseignement à distance, il existe trois approches pédagogiques : cognitivo-comportementale (ou thérapies cognitivo-comportementales, TCC), socio-constructiviste et connectiviste (Anderson et Dron, 2011). Ses approches ont joué un rôle important dans l'enseignement à distance, et ont évolué parallèlement à la technologie. L'apprentissage des langues à distance a également suivi l'évolution. Une évolution qui cherche les moyens les plus efficaces pour une interaction significative. Michael Moore a identifié trois composantes selon le modèle « Trois types d'interaction » pour une interaction de qualité dans un contexte éducatif : interaction apprenant-contenu, apprenant-tuteur et apprenant-apprenant (Moore, 1989). Dans ce cadre, un environnement hautement interactif est un élément clé pour un environnement d'apprentissage efficace. La théorie de l'acquisition d'une langue indique que non seulement les apprenants ont besoin d'un apport compréhensible (Krashen, 1985) (activités de lecture et d'écoute), mais aussi d'occasions de production (Swain, 1995)

(activités pour pratiquer leurs aptitudes discursives, orales et écrites). Les apprenants doivent avoir la possibilité d'interagir avec la langue pour comprendre le sens, rendre les commentaires plus compréhensibles, obtenir des commentaires et reconnaître le besoin de changer leur langue pour réussir la communication. Aujourd'hui avec l'avènement des technologies d'information et de communication, la conception et la mise en place des outils et plateformes d'apprentissage présentent un grand défi pour les professeurs, les concepteurs de cours et les développeurs. Bien sûr, avant de concevoir un cours, ils devraient prendre en considération plusieurs facteurs importants (pédagogie, outils, accessibilité, etc.).

De nombreuses recherches scientifiques ont été menées ces dernières années dans le domaine de l'apprentissage des langues en général (Sokolik, 2014 ; Colpaert, 2014 ; Bárkányi, 2018), et plus précisément, sur l'apprentissage collaboratif assisté par ordinateur qui a montré que les nouvelles technologies peuvent améliorer l'apprentissage des langues. Basé sur certaines théories présentées dans cet article, un environnement d'apprentissage de langue réussi devrait être interactif. Il devrait proposer des possibilités de communication authentique, ainsi qu'un accès au matériel éducatif interactif et motivant pour les apprenants. De cette manière, ils pourront mettre en pratique une véritable communication, s'informer sur la culture de la langue, acquérir et pratiquer toutes les compétences linguistiques de base.

Élaborer un cours et une plateforme d'apprentissage des langues LMOOC est un processus très compliqué. En outre des points déjà analysés, il existe d'autres facteurs à considérer tels que les besoins des apprenants, le degré et le mode d'interaction, le niveau de collaboration ou d'apprentissage et les types d'évaluation. En plus de l'infrastructure technologique disponible, la facilité d'utilisation de l'environnement en ligne, le nombre de participants, le temps et le coût. Tous ses facteurs doivent être traités avec soin.

La question la plus importante qui se pose est de savoir quels sont les éléments de base d'un cours dans un MOOC interactif. En d'autres termes, lorsqu'un cours est ouvert à un grand nombre d'apprenants pour apprendre et pratique une langue, quels sont les éléments de base qui doivent être pris en considération lors de l'analyse et de la conception ?

Ces éléments peuvent être divisés en six catégories (Grover, 2013):

1- **Le contenu :** ressources pédagogiques authentiques; utilisation de multimédia; Variété d'activités qui favorisent toutes les compétences linguistiques et soutiennent la sensibilisation culturelle.

2. **La pédagogie :** Communication (apprenant-apprenant, apprenant-tuteur, communauté de classe ouverte); collaboration (CL) (projets de groupe, forums etc.); intelligence collective; autonomie (autonome/auto-rythmé/apprentissage/réflexion); engagement-motivation; apprentissage basé sur le jeu; nombre d'enseignants.

3. **L'évaluation** : Évaluation / (évaluation ouvert et automatisée par pairs, ou par enseignants), amélioration fondée sur les preuves (analyse et traitement de données); commentaires (commentaires, critiques).

4. **La communauté :** Construire la communauté sociale (médias sociaux - outils d'intégration et d'autres outils technologiques).

5. **L'infrastructure technique :** Nombre maximum de participants, performance de la plateforme, sécurité, utilisabilité.

6. **Questions financières :** Frais de cours ou de certification/accréditation.

Chaque élément doit être étudié avec beaucoup de soin par des équipes pédagogiques, des concepteurs, des enseignants ou des développeurs intéressés par la conception et l'évaluation d'un MOOC.

4. MOOC-IRCAM pour les adultes

Les MOOC-IRCAM pour adultes - comme les autres MOOC hébergés dans le monde par des institutions et établissements - sont conçus selon des principes socioconstructivistes qui suivent le modèle xMOOC décrit ci-dessus, où l'enseignement est intégré et fait partie de l'équipe de conception du cours pour permettre aux apprenants de progresser de manière autonome et indépendamment. L'apprentissage dans ce cas est facilité par une présentation des ressources d'apprentissage bien organisée et structurée ainsi que des activités créées pour atteindre de bons résultats. Le MOOC est centré sur l'apprenant, qui offre un haut degré de flexibilité et contrairement aux autres xMOOC qui ne soutiennent pas l'apprentissage collaboratif (Staubitz *et al.*, 2015), il cherche à encourager l'interaction entre apprenants et tuteurs à travers des outils de discussion où l'apprentissage collaboratif peut se placer en utilisant le dialogue, l'échange entre pairs et la rétroaction, ainsi que les recommandations des tuteurs.

Le MOOC-IRCAM est divisé en six compétences, chacune comporte jusqu'à 14 activités appelées « étapes ». Les compétences de lecture, de prononciation et d'écoute sont développées et mises en pratique à travers des activités de compréhension. Il y a une variété d'activités telles que des quiz, des articles, formations, vidéo et des forums de discussions. Les activités sont créées pour encourager l'utilisation de la langue amazighe. Les discussions suivent de nombreuses activités, offrant l'occasion aux apprenants de consolider ou de réfléchir leur apprentissage. Les discussions sont intégrées dans le contenu d'apprentissage et peuvent être divisées en deux types principaux : ceux qui demandent aux apprenants d'écrire et de poster quelques productions en amazighe et ceux qui demandent aux apprenants de réfléchir ou de commenter un aspect de la culture dans leur propre pays d'origine ou de résidence. Les apprenants ont la possibilité de passer un test de progrès à la fin de chaque compétence et un score est donné.

Dans notre MOOC le contenu est semi-structuré; tout en laissant une certaine flexibilité dans la façon dont les apprenants interagir avec le matériel, il suit une progression claire, en allant du plus simple au plus complexe. Pourtant, qu'on les apprenants naviguent sur la plateforme de façon autonome, ils peuvent compléter les activités dans n'importe quel ordre, organisant ainsi leur propre apprentissage en fonction de leurs intérêts, capacités, besoins, etc.

Comme déjà indiquée, la principale fonctionnalité de collaboration offerte par la plate-forme MOOC-IRCAM est un forum de discussion. Les apprenants peuvent se connecter les uns aux autres, collaborer et partager leurs connaissances. Dans le cours d'amazighe pour adultes, les pratiques de collaboration offertes par l'outil de discussion améliorent l'apprentissage et offrent des activités qui favorisent le partage des connaissances.

Les pratiques de collaboration sont proposées pour aider les apprenants à créer un sentiment de communauté. Chaque discussion est déclenchée par une activité ou un article écrit par l'équipe pédagogique. Ceux-ci peuvent être utilisés pour encourager la pratique de la langue en utilisant la langue amazighe ou développer la conscience interculturelle par la comparaison et la réflexion.

Les apprenants utilisent la langue amazighe pour avoir des conversations simples et significatives avec les autres apprenants et l'équipe pédagogique. De cette manière, les apprenants pratiquent leur langue en travaillant ensemble avec les autres et ceci constitue la base de la formation à distance. Le nombre de commentaires après les sujets de discussion permet aux apprenants d'échanger des informations sur leur propre vie, par exemple ; les repas, leur famille, leur lieu de travail, etc.

L'outil de discussion est également utilisé pour soutenir la conversation autour de sujets concernant la langue et la culture amazighe, qui offre aux apprenants l'occasion d'échanger des connaissances.

Le rôle joué par la communauté d'apprenants est crucial dans un cours ouvert et massif où le soutien et la rétroaction représentent un grand défi pour les tuteurs. En même temps, les apprenants viennent avec une expertise et des connaissances diverses qui peuvent être extrêmement utiles au sein de la communauté. Nous avons intégré le principe de « j'aime » pour récompenser les apprenants qui ont commenté aux autres ou leurs contributions, et les apprenants qui ont volontairement fourni des corrections et des commentaires à leurs pairs. L'identification, le suivi et l'encouragement des participants les plus actifs sont un moyen très efficace pour maintenir leur engagement. De cette manière, une variété de réseaux se forme au sein de la communauté.

5. Discussion sur MOOC-IRCAM

La conception et le développement d'un MOOC présente un certain nombre d'opportunités et de défis. Nous essayons ici d'aborder quelques-uns.

Il existe plusieurs MOOC basés sur une approche connectiviste suivant un modèle largement xMOOC (Beayen *et al.*, 2013). Ce modèle xMOOC est souvent préféré par la plupart des établissements, car il donne l'opportunité de réorienter et innover le contenu du cours sans changer leur culture ou leur approche pédagogique (Moreira *et al.*, 2014).

Un défi important est la possibilité d'avoir un contrôle global sur toutes les discussions, car le nombre de participants à une formation dans un MOOC est extrêmement élevé. Ces discussions et commentaires dépassant parfois les 40 000. La majorité des plateformes n'autorise pas le tri avancé par mot clé, ce qui complique le filtrage et l'analyse des commentaires et discussions, mais permet aux participants de trier par catégorie d'utilisateur soit des apprenants ou tuteurs en sélectionnant uniquement les commentaires et discussions. Il permet également de trier par « j'aime » afin que les commentaires les mieux notés peuvent être facilement localisés.

Les problèmes culturels à cause de l'hétérogénéité des apprenants, les différents antécédents culturels et les expériences d'apprentissage passées, peuvent parfois présenter un ensemble

des opportunités et défis. La diversité culturelle, religieuse, politique et géographique chez les apprenants et les tuteurs, exige un certain niveau de sensibilisation interculturelle en ce qui concerne les questions socioculturelles.

Un autre défi est lorsque le nombre des apprenants qui terminent ou qui participent à un MOOC est faible. Dans ses études Jordan (Jordan, 2015) a constaté que le taux de participation moyen dans ces MOOCs est d'environ 12%. Il y a plusieurs raisons à cela. Premièrement, plusieurs apprenants ne prennent pas la formation au sérieux et ne veulent même pas un certificat de réussite ou d'un document similaire. Deuxièmes, les cours sont gratuits, donc il n'y a pas d'engament financier impliqué, qui ne motive pas les apprenants pour suivirent la formation.

6. Conclusion

Dans cet article, les auteurs ont commencé par un état d'art sur les MOOCs, ou ils ont exposé ces différentes caractéristiques. Par la suite, ils ont tenté d'exposer les étapes clés qui pourraient aider les concepteurs pédagogiques pour la création d'un environnement d'apprentissage interactif et efficace.

Aujourd'hui les MOOC peuvent jouer un rôle important en rendant les ressources éducatives librement accessibles à un public plus large, encouragent l'innovation dans les approches pédagogiques et permettant aux établissements de tester de nouvelles méthodes de partage de cours.

Les MOOC peuvent potentiellement jouer un rôle important en comblant le canal entre l'apprentissage formel et informel et en élargissant la participation. Ils remplissent le mandat de rendre des ressources éducatives librement accessibles à un plus large public, et ils encouragent l'innovation dans les approches pédagogiques, permettant aux universités de tester de nouvelles façons de partager des cours.

L'article présente quelques défis et opportunités pour la conception et la réalisation d'un MOOC pour l'apprentissage de la langue amazighe pour les adultes. En raison de leur échelle massive, les équipes pédagogiques et concepteurs de cours présentent des défis dans la gestion du processus d'apprentissage.

Nous prévoyant de mètre des expérimentations dans des situations réales de formation pour avoir un aperçu de l'expérience de l'apprenant et recueillir des données sur leur participation.

Les plateformes MOOC évolueront de manière exponentielle, deviendront plus sophistiquées, la nature de l'expérience d'apprentissage changera dans le temps et les universités et établissements devront suivent cette évolution.

Références

Anderson, T., Dron, J. (2011). Three Generations of Distance Education Pedagogy, International Review of Research in Open and Distance Learning, Volume 12, Number 3.

Bárcena, E., Read, T., Martín-Monje, E., Castrillo, M. D. (2014). Analysing student participation in foreign language MOOCs: a case study. European MOOCs Stakeholders Summit, pp. 11-17.

Bárkányi, Z. (2018). Can you teach me to speak? Oral practice and anxiety in a language MOOC. Innovative language teaching and learning at university: integrating informal learning into formal language education, Vol. 9.

Beaven, T., Comas-Quinn, A., Hauck, M., de los Arcos, B., Lewis, T. (2013). The open translation MOOC: creating online communities to transcend linguistic barriers. In OER 13 Creating a virtuous circle, Nottingham.

Colpaert, J. (2014). 10 Conclusion. Reflections on Present and Future: towards an Ontological Approach to LMOOCs.

Downes, S. (2011). The role of the educator. The Huffington Post. http://www. huffingtonpost. com/stephen-downes/the-role-of-the educator_b_790 937.html

Downes, S. (2012). Connectivism and connective knowledge: Essays on meaning and learning networks. Stephen Downes w eb.

Downes, S. (2012). Massively open online courses are 'here to stay'. Stephen Downes. http:// www.downes.ca/post/58676

Ferguson, R., Coughlan, T., Heredotou C. (2016). MOOCs: what the Open University research tells us. The Open University, Milton Keynes: Institute of Educational Technology.

Grover S., Franz P., Schneider E., Pea R. (2013). The MOOC as distributed intelligence: Dimensions of a framework & evaluation of MOOCs.

Hurd, S. (2006) Towards a better understanding of the dynamic role of the distance language learner: learner perceptions of personality, motivation, roles and approaches. Distance Education 27 (3): 299-325.

Jordan, K. (2015). Massive open online course completion rates revisited: assessment, length and attrition. International Review of Research in Open and Distributed Learning, 16(3):341–358.

Kennedy, J. (2014). Characteristics of massive open online courses (MOOCs): a research review, 2009-2012. Journal of Interactive Online Learning. Vol. 13, n° 1.

Krashen, S. (1985). The input hypothesis: Issues and implications. London: Longman.

Littlejohn, A. (2013). Understanding massive open online courses. New Delhi: CEMCA.

Mackness, J., Mak, S., w illiams, R. (2010). The ideals and reality of participating in a MOOC. Paper presented at the Seventh International Conference on Networked Learning, Aalborg, Denmark.

McGreal, R., Kinuthia, w ., Marshall, S., McNamara, T. (2013). Open educational resources: Innovation, research and practice. Commonwealth of Learning (COL).

Moore, M.G. (1989). Three types of interaction. The American Journal of Distance Education, 3(2): 1-6.

Moreira Teixeira, A., Mota, J. (2014). A proposal for the methodological design of collaborative language MOOCs. In Martin-Monje, E., Bárcena, E. (Eds), Language MOOCs: providing learning, transcending boundaries. De Gruyter Open.

Pappano, L. (2012). The year of the MOOC. The New York Times. http://www.edinaschools.org/cms/lib07/MN01909547/Centricity/Domain/272/The%20Year%20of%20the%20MOOC%20NY%20Times.pdf

Parr, C. (2013). MOOC creators criticise courses' lack of creativity. Times Higher Education. https://www.timeshighereducation.com/news/mooc-creators-criticise-courses-lack-of-creativity/2008180.

Rodriguez, O. (2013). The concept of openness behind c and x-MOOCs (Massive Open Online Courses) Open Praxis, 5(1):67-73. Retrieved from http://dx.doi.org/10.5944/openpraxis.5.1.42

Rodriguez, O. (2012). MOOCs and the AI-Stanford like Courses: two successful and distinct course formats for massive open online courses. European Journal of Open, Distance, and E-Learning, July 5th, 2012. Retrieved from http://www.eurodl.org/?article=516

Siemens, G. (2013). Massive open online courses: innovation in education? In R. McGreal. Kinuthia, w ., Marshall S. (Eds), Open educational resources: innovation, research and practice. pp. 5-16. Commonwealth of Learning, Athabasca University.

Sokolik, M. (2014). what constitutes an effective language MOOC. Language MOOCs: Providing learning, transcending boundaries. pp. 16-32.

Staubitz, T., Pfeiffer, T., Renz, J., w illems, C., Meinel, C. (2015). Collaborative learning. In ICERI2015 8th annual International Conference of Education, Research and Innovation, Seville, Spain. pp. 18-20.

Swain, M. (1995). Three functions of output in second language learning. In Cook, G., Seidlhofer, B. (Eds.), For H.G. w iddowson: Principles and practice in the study of language. pp. 125–144. Oxford: Oxford University Press.

Yuan, L., Powell, St. (2013). MOOCs and Open Education: Implications for Higher Education. A white paper, CETIS. Retrieved from http://publications.cetis.ac.uk/2013/667

TICAM'18, p. 21 - 27

Promouvoir l'enseignement du codage informatique chez les enseignants de la langue amazighe au primaire marocain

Ibrahim OUAHBI[1,2], Hassane DARHMAOUI[2], Fatiha KADDARI[1], Mohamed REGRAGUI[3]

[1] Université Sidi Mohamed Ben Abdellah, - FSDM – Laboratoire de Didactique et d'Innovation Pédagogique et Curriculums -Fès, Maroc

[2] Center for Learning Technologies (CLT), Université Al Akhawayn , Ifrane, Maroc

[3] Centre Régional des Métiers d'Education et de Formation de l'Oriental (CRMEF), annexe provinciale de Nador, Maroc

ibrahim.ouahbi@usmba.ac.ma h.darhmaoui@aui.ma

kaddari@yahoo.fr reg.mohamed@gmail.com

Résumé

Notre objectif est de contribuer à l'innovation pédagogique dans l'enseignement de la langue amazighe grâce à l'intégration de la pensée algorithmique dès l'enseignement primaire. Pour ce, nous avons mené avec 26 enseignants stagiaires de la langue amazighe ; à l'annexe provinciale du Centre Régional des Métiers de l'Education et de la Formation (CRMEF) de l'oriental de la ville Nador ; des activités ludiques utilisant l'environnement visuel de programmation « Scratch ». Au cours de ces séances les enseignants stagiaires ont été initiés à la logique algorithmique et à la programmation visuelle à base de blocs, après ils ont été invités à créer des programmes éducatifs avec l'outil Scratch. En amont et en aval de notre expérimentation un questionnaire a été distribué aux enseignants stagiaires participants pour comparer l'évolution de leurs perceptions sur les activités menées. L'analyse des résultats a montré une bonne prise de conscience sur les différentes possibilités d'exploiter l'environnement Scratch afin de faciliter l'enseignement et l'apprentissage de la langue amazighe tout en développant la pensée algorithmique chez les apprenants dès leurs jeunes âges.

Mots clés : Environnement Scratch, Technologies d'Information et de Communication en Education (TICE), Formation des enseignants, Langue amazighe, Informatique au primaire.

1. Introduction

L'informatique est considérée actuellement comme une composante fondamentale des cursus éducatifs et l'enseignement du codage informatique est une tendance croissante à l'échelle internationale (Ouahbi *et al.*, 2015b ; Dagiene *et al.*, 2016 ; Fluck *et al.*, 2016 ; Kanemune *et al.*, 2017 ; webb *et al.*, 2017 ; Arcos *et al.*, 2018). Dans ce sens, pour promouvoir l'enseignement de la science informatique et en particulier la pensée algorithmique chez les élèves du primaire, il est important de fournir aux enseignants une connaissance adéquate de la science informatique et de voir comment l'intégrer dans leur enseignement.

Ainsi, dans cet article, nous tentons répondre à la question de recherche suivante : «*l'intégration des activités TICE basées sur le codage informatique à l'aide d'un environnement visuel de programmation dans la formation des enseignants, entraine-elle des attitudes positives de ces futurs enseignants envers le codage informatique* ?»

Pour ce, nous avons dans un premier temps dressé un état de l'art des travaux rattachés à l'enseignement de l'informatique au primaire et à l'étude de la perception des enseignants envers le codage informatique. Ensuite, nous avons réalisé pendant 3 séances (2 heures / séance) des activités d'introduction à la logique algorithmique et à la création de programmes à l'aide de l'environnement Scratch[1] (Maloney *et al.*, 2010) avec 26 enseignants stagiaires de la filière «langue amazighe» au CRMEF de l'oriental, annexe provinciale de la ville Nador. Ces séances ont été intégrées au module TICE, avec la présence du formateur chargé d'enseigner ce module. En amont et en aval de l'expérimentation, un questionnaire a été distribué à ces enseignants stagiaires pour mesurer leurs perceptions sur le codage informatique au primaire. L'étude comparative de l'évolution de leur perception a montré une très bonne conscience sur le potentiel du codage informatique dans le développement de la pensée algorithmique, la communication en langue amazighe ainsi que d'autres compétences transversales.

2. Etat de l'art

2.1. L'informatique au primaire

Parmi les finalités de l'enseignement au primaire, on peut citer celle de rendre l'élève capable d'utiliser les nouvelles Technologies d'Information et de Communication (TIC) (MEN, 2011). Cependant, en analysant les instructions officielles on constate que le terme informatique est cité juste une seule fois. Ces instructions ministérielles présentent, mais d'une façon ambigüe, des indications sur l'utilisation des ressources numériques dans l'enseignement au primaire.

Dans la pratique, l'exploitation des TIC reste très restreinte et limitée à des initiatives d'une minorité d'enseignants (El ouidadi *et al.*, 2013). On peut aussi souligner quelques expériences de l'enseignement de l'informatique dans le secteur privé, où l'informatique est intégrée pour attirer plus d'élèves. Cependant, quelques manuels réalisés portent sur des activités ludiques pour le développement de la pensée algorithmique chez l'apprenant (Bemmouna *et al.*, 2013).

2.2. Formation des enseignants et informatique

Dans les dispositifs de la formation aux CRMEF (MEN, 2012), actualisés en 2014, les enseignants stagiaires sont initiés à l'informatique, outil via le module TICE. Les notions de la pensée algorithmique et de la science informatique ne sont abordées que pour les filières de la spécialité informatique. Parmi les objectifs des curricula relatifs aux modules de formation des futurs enseignants aux CRMEF en matière de TICE, on peut citer les compétences suivantes :

- être capable de travailler dans un environnement numérique.

1 https://scratch.mit.edu/

- être capable de produire, traiter, exploiter et diffuser des documents numériques.
- être capable de s'informer et se documenter par le biais des TIC.
- être capable de communiquer et d'échanger les données avec les TIC.
- être capable de choisir une ressource numérique à intégrer.
- être capable de gérer une séance intégrant des ressources numériques.
- …

Il parait donc que l'informatique est utilisée en tant qu'outil et les notions de base de la science informatique sont presque absentes dans les curricula de la formation des enseignants aux CRMEF. Pourtant, il nous semble qu'il est important de fournir aux enseignants une connaissance adéquate des fondamentaux de la pensée algorithmique et du codage informatique, pour ainsi promouvoir l'enseignement de l'informatique science dès le primaire.

2.3. Le codage informatique dans la formation des enseignants

Plusieurs recherches se sont intéressées à l'étude des attitudes des enseignants après leurs expositions à des séances de sensibilisation sur le codage informatique. Romero *et al.* (2015), avec 51 enseignants stagiaires, ont révélé qu'avant l'expérimentation ces participants ignoraient les possibilités de réutiliser des programmes existants ou en créer de nouveaux en utilisant des environnements de programmation visuelle à base de blocs. Ces données vont dans la même lignée avec les résultats de notre expérimentation menée avec 20 enseignants stagiaires du cycle primaire spécialité «langue amazighe» (Ouahbi *et al.*, 2016), à savoir que la totalité de ces futurs enseignants, ignoraient les possibilités de création de programmes. Bien que la personnalisation et la création de programmes demandent des compétences informatiques avancées (Earp *et al.*, 2013), il existe des environnements informatiques qui facilitent la programmation aux débutants à travers la création des jeux, d'animations et d'histoires (Ouahbi *et al.*, 2015a). Dans ce contexte plusieurs travaux ont prouvé l'efficacité de ces environnements, en particulier l'environnement Scratch, dans l'enseignement non seulement des concepts de base de la programmation, mais aussi l'acquisition des compétences transversales comme la communication, le partage, la pensée algorithmique, la résolution de problèmes, l'autonomie et la créativité chez les élèves (Ouahbi, 2018). L'environnement Scratch a déjà fait ses preuves au niveau universitaire. Kim *et al.* (2012) l'ont utilisé dans des cours de programmation au profit d'enseignants en cours de formation. Ces derniers ont créé des animations, des histoires et des jeux pour des fins éducatifs. Ils ont développé des compétences en programmation d'une façon implicite et ils ont renforcé leurs capacités en TIC à fin d'innover et d'intégrer ces nouvelles technologies dans leurs pratiques enseignantes. De même, Baron *et al.* (2013) ont mené une expérimentation d'initiation à la programmation en utilisant l'environnement Scratch auprès d'étudiants de Master en sciences de l'éducation dans le cadre d'une pédagogie par projet. Les résultats de cette étude ont montré qu'il est possible, de faire produire des programmes relativement intéressants par des personnes non formées auparavant en informatique.

3. Méthodologie

Nous avons mené avec 26 enseignants stagiaires du cycle primaire spécialité «langue amazighe» des activités ludiques utilisant l'environnement visuel de programmation Scratch. Tous les participants à cette expérimentation suivent leur formation au CRMEF de l'orientale, annexe provinciale de Nador.

Nous avons commencé notre expérimentation par l'identification de l'expérience et de l'avis de ces enseignants stagiaires sur l'informatique via un questionnaire. Ensuite, ces participants ont suivi une formation de 6 heures étalées sur 3 séances.

Durant la 1ère séance, les notions de base de la logique algorithmique et de la programmation ont été introduites à travers des activités ludiques : les labyrinthes Blockly[2] et des ressources disponibles sur code.org[3]. Lors de la 2ème séance, les enseignants stagiaires ont été initiés à l'environnement Scratch en réalisant des animations et des petits jeux accessibles à partir du menu «Aide» du logiciel. Au cours de la 3ème séance les enseignants stagiaires ont été amenés à créer des programmes, des histoires et des dialogues en langue amazighe avec Scratch. L'objectif était de les initier à créer leurs propres programmes éducatifs. A la fin de cette séance nous leurs avons proposé de réaliser des fiches pédagogiques d'une leçon qui porte sur le codage informatique en relation avec l'enseignement de la langue amazighe à l'école primaire.

Nous soulignions que pendant les 3 séances, nous avons échangé des discussions avec les stagiaires sur les différentes manières et approches pour l'enseignement de la logique algorithmique et une meilleure exploitation de l'environnement Scratch en cours de la langue amazighe.

A la fin de ces séances TICE, les enseignants stagiaires ont répondu à un questionnaire pour identifier leurs attitudes et avis sur le codage informatique et l'apprentissage de la langue amazighe au primaire à l'aide de l'environnement Scratch. A noter que les mêmes enseignants ont répondu à un questionnaire avant notre expérimentation. L'objectif était d'étudier l'évolution de leur perception vis-à-vis la programmation intégrée dans l'enseignement de la langue amazighe.

4. Résultats et discussions

La quasi-totalité des enseignants stagiaires participants (24 parmi 26) à notre enquête ont suivi des cours d'informatique pour au moins une année pendant leur parcours scolaire (10 ont suivi ces cours au secondaire collégial, 9 au secondaire qualifiant et 7 lors de leurs études supérieures). La majorité des participants qui ont suivi des cours d'informatique durant leur parcours scolaire semble être insatisfaite de leurs savoirs relatifs à la création des programmes informatiques, la production des documents multimédias, et la création des pages wEB. La majorité aussi affirme avoir acquis les notions de base d'un logiciel de traitement de texte et que les cours d'informatique qu'ils ont suivi leur a permis de développer des compétences de manipulation de tableurs et de présentation.

2 https://blockly-games.appspot.com/?lang=fr
3 https://studio.code.org/

Lorsque les participants ont été questionnés sur leur expérience avec l'algorithmique et la programmation : presque la totalité (23 parmi 26) semble ignorer le terme «algorithme» et seul 8 d'entre eux ont déclaré avoir assisté à un cours de programmation durant leur parcours scolaire. Ces résultats confirment ceux obtenues lors d'une étude similaire avec les élèves du secondaire collégial et secondaire qualifiant à savoir qu'un taux important d'apprenants éprouvent des attitudes négatives sur leurs apports en discipline informatique en particulier la programmation, la production des documents multimédias et la création des pages wEB (Ouahbi, 2018).

Les résultats de l'étude de l'évolution de la perception des 26 enseignants stagiaires sur le codage informatique comme objet d'enseignement au primaire sont résumés dans le tableau suivant :

Question ***l'enseignement de la programmation en classe pourrait :***	**Avant**	**Après**
a. aider les élèves à développer leur imagination	19	26
b. aider les élèves à apprendre à coopérer avec les autres	15	24
c. améliorer la capacité de la résolution de problèmes chez les élèves	13	23
d. développer des compétences en communication chez les élèves	2	26
e. motiver les élèves pour l'apprentissage de la langue amazighe	1	26
f. aider les élèves, en général, dans leurs apprentissages au primaire	1	26

Tableau 1: Étude comparative de l'évolution de la perception sur la pensée algorithmique avant et après l'expérimentation

Il ressort de la lecture comparative des données ci-dessus que les enseignants stagiaires, une fois initié à la pensée algorithmique et à l'outil Scratch, ont complétement changé leurs attitudes vis-à-vis l'intégration implicite de la programmation dans le mode d'enseignement, pour motiver les apprenants non seulement pour la langue amazighe et communication, mais aussi tous les apprentissages au primaire. En effet, le nombre des avis favorables est passé de moins de 2 en amont de l'expérimentation à 26. Après l'expérimentation, le nombre des enseignants stagiaires qui pensent que l'intégration du codage informatique en classe pourrait développer l'imagination des élèves a passé de 19 à 26. Ceux qui pensent que le codage informatique peut aider à développer des compétences de collaboration est passé de 15 à 24. Le nombre des favorables à la proposition «*améliorer la capacité de la résolution de problèmes chez les élèves* » a évolué de 13 à 23.

Les séances menées ont donné l'occasion aux 26 enseignants stagiaires, du cycle primaire spécialité «langue amazighe» du CRMEF de l'orientale annexe provinciale de Nador, qui ont participé à notre expérimentation d'apprécier l'apport positif que peut apporter

l'enseignement de la logique algorithmique via l'environnement Scratch et ils ont pu saisir le potentiel éducatif de cet environnement en primaire et en particulier dans l'enseignement et l'apprentissage de la langue amazighe. A la fin, ces participants ont recommandé l'intégration d'un environnement visuel de programmation à base de blocs dans la formation initiale aux CRMEF ou dans la formation continue des enseignants.

5. Conclusion

Les résultats de notre travail ont indiqué que la plupart des futurs enseignants participants doutaient du potentiel et de l'apport positif de l'introduction du codage informatique en classe et en particulier dans l'enseignement de la langue amazighe. Pourtant, une fois qu'ils ont été initiés, leur perception a considérablement évolué. En effet, en aval de l'expérimentation la totalité des participants semble être favorable pour le rôle du codage informatique dans la motivation, le développement de compétences en communication en langue amazighe chez les élèves. Les enseignants stagiaires ont apprécié l'environnement Scratch et ont saisi son potentiel éducatif. Ils recommandent l'intégration d'un tel environnement facilitant la création vidéo-ludique dans la formation initiale aux CRMEF ou dans la formation continue des enseignants.

Références

Arcos, G., Aguirre, G. L., Hidalgo, B., Rosero, R. H., Gómez, O. S. (2018). Current Trends of Teaching Computer Programming in Undergraduate CS Programs: A Survey from Ecuadorian Universities. *KnE Engineering, 1(2)*:253-275.

Baron, G. L., Voulgre, E. (2013). Initier à la programmation des étudiants de master de sciences de l'éducation? Un compte rendu d'expérience. In Sciences et technologies de linformation et de la communication (STIC) en milieu éducatif.

Bemmouna, A., Chetouani, A., El yacoubi, T. (2013). *L'informatique pas à pas. Collection de l'enseignement de l'informatique au primaire (du 3ème au 6ème)*. Rabat : Dar Nachr Almaarifa.

Dagiene, V., Stupuriene, G. (2016). Informatics concepts and computational thinking in K-12 education: A Lithuanian perspective. *Journal of Information Processing, 24(4):* 732-739.

Earp, J., Dagnino, F., Kiili, K., Kiili, C., Tuomi, P., Whitton, N. (2013). Learner Collaboration in Digital Game Making: An Emerging Trend. Learning & Teaching with Media & Technology, 439.

El ouidadi, O., Lakdim, A., Essafi, K., Sendide, K., Depiereux, E. (2013). Principaux facteurs influençant les usages des TICE chez des enseignants marocains. *Frantice.net*, Vol.6, pp. 37-52.

Fluck, A., Webb, M., Cox, M., Angeli, C., Malyn-Smith, J., Voogt, J., Zagami, J. (2016). Arguing for computer science in the school curriculum. *Journal of Educational Technology & Society, 19(3):*38.

Kanemune, S., Shirai, S., Tani, S. (2017). Informatics and Programming Education at Primary and Secondary Schools in Japan, Olympiads in Informatics, Vol. 11, pp.143–150.

Kim, H., Choi, H., Han, J., So, H. J. (2012). Enhancing teachers' ICT capacity for the 21st century learning environment: Three cases of teacher education in Korea. Australasian Journal of Educational Technology, 28(6):965-982.

Maloney, J., Resnick, M., Rusk, N., Silverman, B., Eastmond, E. (2010). The scratch programming language and environment. ACM Transactions on Computing Education (TOCE), 10(4):16.

MEN. (2011). Ministère d'Education Nationale, Programme et instructions officielles pour l'enseignement au cycle primaire.

MEN. (2012). Ministère d'Education Nationale, Guide des modules transversaux aux CRMEF, unité centrale de la formation des cadres.

Ouahbi, I., Darhmaoui, H., Kaddari, F., Elachqar, A., Lahmine, S. (2015a). Vers un enseignement de la programmation compatible avec la culture vidéoludique des élèves au Maroc. Actes de la 7ème Conférence sur les Environnements Informatiques d'Apprentissage Humain EIAH'15, Agadir, Maroc

Ouahbi, I., Darhmaoui, H., Kaddari, F., Bemmouna, A., Elachqar, A., Lahmine, S. (2015b). Un aperçu sur l'enseignement de l'informatique au Maroc: Nécessité d'une réforme des curricula - An overview of teaching informatics in Morocco: The need for a curriculum reform. *Frantice.net*, Vol.11, pp.51-66.

Ouahbi, I., Darhmaoui, H., Kaddari, F. (2016). Perception évolutive envers les Serious Games et à la création vidéoludique par des enseignants stagiaires de la langue amazighe. 7ème édition de la conférence internationale sur les Technologies d'Information et de Communication pour l'Amazighe, Rabat, Maroc.

Ouahbi, I. (2018). Serious games : exemple d'innovation pédagogique en classe d'informatique au secondaire marocain (PhD Thesis). Fès, Maroc: Université sidi Mohammed ben Abdellah.

Romero, M., Barma, S. (2015). Teaching pre-service teachers to integrate Serious Games in the primary education curriculum. *International Journal of Serious Games*, Vol.2, n° 1.

Webb, M., Davis, N., Bell, T., Katz, Y. J., Reynolds, N., Chambers, D. P., Sysło, M. M. (2017). Computer science in K-12 school curricula of the 2lst century: Why, what and when?. *Education and Information Technologies, 22(2):*445-468.

TICAM'18, p. 29 - 45

Enhancing the learning experience with pop-up feature in flying dictionary Android application

Deeheem ANSARI, Gurtej KOCHAR
Netaji Subhas Institute of Technology, New Delhi, India
{deeheem_ansari, gurtej_kochar}@yahoo.in

Abstract

We develop an Android application Flying Dictionary which is an offline pop-up dictionary meant to ease users by displaying the meaning of the words they find tough without having to switch to some other app in order to look for its meaning. The application was created and uploaded on Play Store in March 2017, and since then over 4000 users have installed it on their Android devices. w ith a current rating of 4.8 (as of December 2017), our aim is to make it easier for otherwise reluctant students to encourage learning new words with ease and at the click of a button. This paper analyses the impact of using the application for daily look-up of meanings, on users' learning experience. The analysis indicates that providing the users with the ability to glance the meanings on the go helps them in learning more new words and prevents them from giving excuses like the need for an online connection or the need for switching the app in order to look for the meaning. This contributes to the growing importance of the pop-up feature and emphasizes the need to incorporate this in the popular dictionary apps as well as try and bring up the feature as a built-in utility in the Android OS itself.

Keywords : e-learning, Dictionnary, Pop up features, Android application.

1. Introduction

By the middle of the seventeenth century English Language had more or less assumed its present form so far as grammar, spelling and pronunciation are concerned. The chief developments have been in the direction of an enlargement of the vocabulary on the one hand and changes in the meanings of words on the other. Hence it can be stated that: as knowledge grows, so does the language with it. The English language is one of the richest and has the most extensive vocabulary (Rajarajeswari andMohana,, 2013). The Oxford dictionary editors themselves are quite modest about this. They write "On an average, we add approximately 1,000 new entries to Oxford Dictionaries online every year" (Manivannan, 2015). In this paper, we propose and evaluate the need for a dictionary that is (i) offline, i.e. does not require net connectivity and, (ii) equipped with a pop-up facility, that is, it helps the user get the meaning of any and every word and many common phrases at just the selection of a word. w e chose the Android platform for implementing this dictionary as it is most widely used (Farkade and . Kaware, 2015), but the core concept can be ported to any other platform.

1.1. Importance of Vocabulary

As English is a second language in most of the countries (Shaunacy, F., 2017), vocabulary knowledge is often viewed as a critical tool for second language learners because a limited vocabulary in a second language impedes successful communication. Underscoring the importance of vocabulary acquisition, Schmitt (2000) emphasizes that lexical knowledge is central to communicative competence and to the acquisition of a second language.

1.2. Positive effects of clarifying meanings simultaneously

It is a well-known fact that people tend to skip looking for meanings of difficult words while reading some piece of text, the reasons being numerous like the dictionary is not handy, or they are just too lazy to search for the word. The disadvantage of this whole act is they tend to lose out on their vocabulary. If any means is available by which the meanings are readily available to people at the click of a button, then all these excuses will hold no value. Moreover, it will increase their vocabulary and make them more intellectual.

1.3. How technology has helped in learning vocabulary

Technology has provided users with creative freedom, endless resources and learning materials, and the possibility to learn from resources available from all corners of the globe. However, the increasing trend is towards using mobile devices to connect to the web. Mobile learning (or m-learning) is the ability to learn anywhere and at any time using a portable electronic device.

1.4. Importance of Android

Figure 1 showing the explosive growth of Google's Android operating system over the last several years (Patange, 2016):

Enhancing the Learning Experience with Pop-Up Feature in Flying Dictionary Android Application

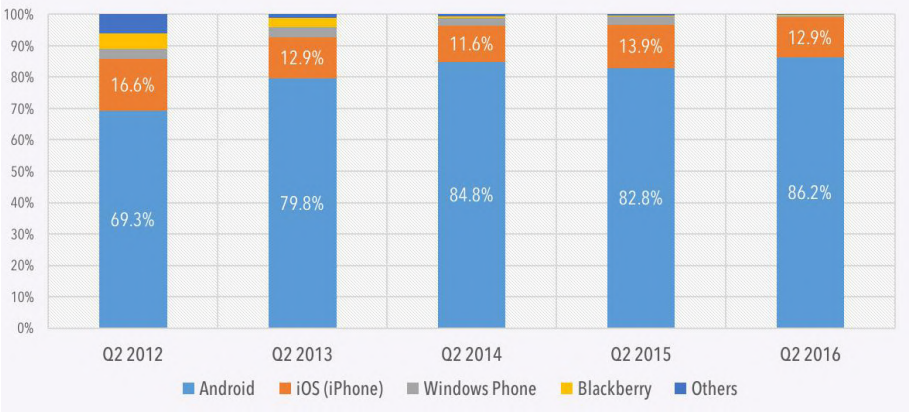

Figure 1: Smartphone OS Market Share (Source: IDC Tracker, Gartner, Aug 2016)

Android being a free solid mobile operating system proves to be a game changer for the developing world where mobile connectivity is cheaper and more reliable than its wired counterpart. It allows phone manufacturers to produce relatively advanced devices without worrying about software this makes them cheaper and gets them into the hands of more people. Thus using Android as a platform promises to reach out to more people.

2. Related Work and Motivation

One of the most common phrases publicized by Google has been search the web. This one feature that allows users to look up the web for any apt definition/data that is related to the words/ phrases in context. Our first and foremost aim is to solve a cliché issue of having the meanings available in an offline mode so that minimum usage of the internet is made. According to Hsiu-Fen Lin, he perceived ease of use and deduced that offline activities are determinants of sustainability of virtual communities (Lin., 2007). Hence, we use the w ordNet database, which is one such open source database, which made implementing of offline dictionaries possible (Redkar et. al, 2015,). Many of the other offline dictionaries also rely on this database for word-meanings, English Dictionary w ordNet being one of them (DictData, 2014), but not providing with the pop-up feature yet. Google Chrome gives no offline feature, although it has the pop-up feature, but it only works on the app itself, and is not available over other apps or PDFs (Gordon, 2015).

Our first objective is to aid bibliophiles who are averse to go to Google for every tough word that is faced, and to ex-iOS users who miss the notify look-up feature (Filipwicz, 2016). This valuable feature i.e. powered by Apple in their iOS phones has seen vast usage, but is currently not available in Android smartphones (Ozaz, 2014). w e found motivation to work in this area to give Android users the same features they would get in an Apple device. Our second objective is to provide users the convenience of seeing the possible meanings of a touch word as they go, without necessarily having an internet connection. The application we developed thus combines offline features with pop-up of meanings on the fly.

Looking into the cognitive load theory, the possibility of dictionary consultation being a hindrance to reading comprehension or incidental vocabulary learning can be easily explained. The cognitive load theory postulates that there are three types of cognitive loads that have an impact on learning performance:

Extraneous cognitive load: This refers to irrelevant information limiting the capacity of the working memory. This load is generated by the presentation method of instruction, and attributable to the design of the instructional materials and the mode in which new information is presented to learners in an educational setting. An example of extraneous cognitive load occurs when there are two possible ways to present new vocabulary; e.g., using a paper-based, bilingual dictionary or an online pop-up dictionary. In this instance, the efficiency of the pop-up dictionary is preferred as it does not unduly load the learner with unnecessary information, which is extraneous.

Intrinsic cognitive load: This term refers to the basic memory capacity for holding task elements in the working memory for simultaneous cognitive processes. This load is defined by the inherent level of difficulty associated with a specific instructional topic.

Germane cognitive load: This term refers to the efforts on the part of the learner to facilitate learning tasks. It is the load devoted to the processing, construction and automation of schemas. According to several researchers (Liu and Lin, 2011; Paas *et al.*, 2004; Sweller, 2010), the method and format of gloss presentation can either reduce or increase cognitive loads associated with the learning tasks, thereby facilitating or hindering task learning performance. For instance, integrating glossaries into the reading passages (integrated format) can potentially decrease extraneous cognitive loads, increase memory, and enhance relevant cognitive loads; consequently leading to better reading comprehension. In addition, the use of pictorial cues, animation, multimedia, lexical annotations and glosses (or a multimodal method of word presentation) was found to be conducive to enhanced incidental vocabulary acquisition (AbuSeileek, 2011).

The convenience of using digital dictionaries is widely appreciated; of which there can be many different types of e-dictionaries recommended for use, since immediate feedback from e-dictionaries while reading facilitates vocabulary acquisition more favorably (Zhiliang, 2008). Commonly suggested ones are type-in online dictionaries and pop -up electronic dictionaries such as those installed on the w indows operating system. These are readily available for use with a move of a cursor or a double-click on any given word to bring up a definition. The uses of these types of electronic dictionaries have benefits and drawbacks. For example, there are arguments against the casual use of pop-up dictionaries, which are considered to be ineffective for long-term retention of vocabulary. These arguments relate to the effect of cognitive load and effective manipulation of the working memory for rehearsing, recalling and retaining new meanings. (Anderson, 1995; Hulstijn, 2001; Barrouillet *et al.*, 2007)

Although quite a many related apps are available on the app store, however, not much of prior work is available on the subject as yet. But we find an excellent research conducted by Alharbi (Alharbi, 2016) where he compares four types of dictionaries, namely pop -up dictionary, type-in dictionary, book dictionary, and no dictionary aid for improving reading comprehension and vocabulary learning. His experiments indicated that the pop-up dictionary group had the shortest average vocabulary searching time, vocabulary and text reading time, and more look-ups ($p<0.001$) than other dictionary groups. Reading comprehension and vocabulary learning were higher for the pop-up dictionary group than for other dictionary groups. Furthermore, survey data indicated that pop-up dictionary participants had slightly more positive attitudes toward dictionary use than the type-in group, and both had significantly more positive attitudes than book dictionary participants. These findings motivated us even more to work in the direction of pop-up dictionaries and give Android users this lacking feature.

Answers to the following research questions to determine whether similar results would be found from the pop-up Flying Dictionary app will be looked into in further. Data was extrapolated from the users of the application via Google Docs. The results should be relevant for foreign language instruction in any location where the target language is not widely used in the local population.

1. Whether the use of pop-up feature can lead to the most efficient dictionary use, reading comprehension and vocabulary learning?

2. whether the use of pop-up feature can lead to the most positive evaluation by foreign language readers undertaking a reading task?
3. whether the use of pop-up feature can result in better vocabulary learning and better reading comprehension?

3. About Flying Dictionary

Flying Dictionary can be used to get word definitions inside any reading app (eBooks/News/Browsers etc.) without opening the dictionary app from outside. It saves lot of time while reading and makes search for definition very fast. The application displays words with audio pronunciations, definitions and example sentences as well as the ability to share, save in history or favorites to provide users tools for enhancing vocabulary without any formal instructions. It has a very simple GUI that makes it quite easy to use. This tool also employs techniques to provide users with a word of the day.

3.1. Features

- Simple GUI
- Extensive database of words including multiple definitions, synonyms and usages
- word of the Day
- History/Favorite word list
- Text to Speech
- Share word or history/favorites list
- Google search to know more
- Offline & Faster
- Search auto-compete
- Auto-start on booting up of device
- Compatible with Browsers (Chrome, Opera), PDF Reader (Adobe Acrobat), E-book Reader (Moon+ Reader), Facebook, whatsApp and many more

Figure 2: App User Interface

Figure 3: Pop-up User Interface

3.2. Maximum Adaptability, Minimum Redundancy

3.2.1 Supporting Different Devices

Android devices come in many shapes and sizes all around the world. w ith a wide range of device types, it is definitely an opportunity to reach a huge audience with our app. In order to be as successful as possible on Android, an app needs to adapt to various device configurations. Out of which some of the important variations that should definitely be considered include different languages, screen sizes, and versions of the Android platform.

3.2.2 Supporting Different Screens

Android categorizes device screens using two general properties: size and density. Learning from the previous experiences, the app had to be built in a form that it was adaptable on devices with screens that range in both size and density. Hence we have included some alternative resources that optimize the app's appearance. (In the form of different layouts and bitmaps used for different screens). Also, the screens orientation (landscape or portrait) is considered a variation of screen size; hence adaptability of the layout was an important factor to optimize the user experience in each orientation.

3.2.3 Supporting Different Platform Versions

while the latest versions of Android often provide great APIs for apps, one should continue to support older versions of Android until more devices get updated. Flying Dictionary takes advantage of the latest APIs while continuing to support older versions as well. As a matter of fact, the dashboard for Platform Versions is updated regularly to show the distribution of active devices running each version of Android, based on the number of devices that visit Play Store. Thankfully Flying Dictionary was able to support more than 90% of the active devices, while targeting the app to the latest version.

4. Technical Work

Keeping in mind the blueprint of the look up feature of iOS, the aim was to design a similar tool on the Android platform. Android is a mobile operating system developed by Google, based on the Linux kernel and designed primarily for touchscreen mobile devices such as smartphones and tablets. Thus studying development on android platform was one of the major technical aspects we had to look into.

Categorizing into broad categories, the following aspects of Android development platform have been include:

4.1. Android Studio

Android Studio is the official integrated development environment for Google's Android operating system, which is the primary software used in developing Flying Dictionary.

4.2. Intents and Intent Filters

Intents facilitate communication between components in numerous ways.

The three fundamental use cases in our app are:

4.2.1 Starting an activity

Flying Dictionary uses many such activities like MainActivity, LogoActivity, SettingsActivity etc. The way we switch between these activities is using intents.

4.2.2 Starting a service

Intents describe which service to start and carry any necessary data. Flying Dictionary uses services in the form of Clipboard service. It is the primary principle service on which Flying Dictionary works. Once any word is selected and copy is clicked, the corresponding word is read from the clipboard and the meaning is retrieved from the database.

4.3. Delivering a Broadcast

Flying Dictionary is designed to receive a broadcast every time the phone is switched on. The broadcast receiver is responsible for starting the clipboard service again after the phone reboots.

4.4. Notification Manager

Notification manager is used to deliver notifications to a user in the form of Word of the Day, once in 24 hours. In order to let the user be convenient with the timings, the notification time can be reset from settings manager. As per choice, the notifications can also be disabled.

4.5. Fragments

Flying Dictionary uses fragments on its main activity where the user can switch between 3 tabs. These tabs represent three different fragments. The first fragment provides a list of words in lexical order. The second provides a list of favorites, while the third list is history of all the words that the user has looked for through the app.

4.6. Permissions

System permissions are divided into several protection levels. The two most important protection levels are normal and dangerous permissions:

4.6.1 Normal Permissions

Covers areas where the app needs to access data or resources outside the app's sandbox, but where there's very little risk to the user's privacy or the operation of other apps. The system automatically grants such permission to the app. Flying Dictionary uses the time zone permission, to perform time based operations like handling notification time etc.

4.6.2 Dangerous Permissions.

Covers areas where the app wants data or resources that involve the user's private information, or could potentially affect the user's stored data or the operation of other apps. Flying Dictionary uses dangerous permissions in the form of read and write services to the external directory where the database is placed.

5. WordNet Database

w ordNet is an online lexical resource which expresses unique concepts in a language. English WordNet is the first WordNet which was developed at Princeton University. One major reason for using it is that it is a centralized database which can be used in bigger applications easily. Figure 4 shows a simple block diagram of the w ordNet database that is used in Flying Dictionary.

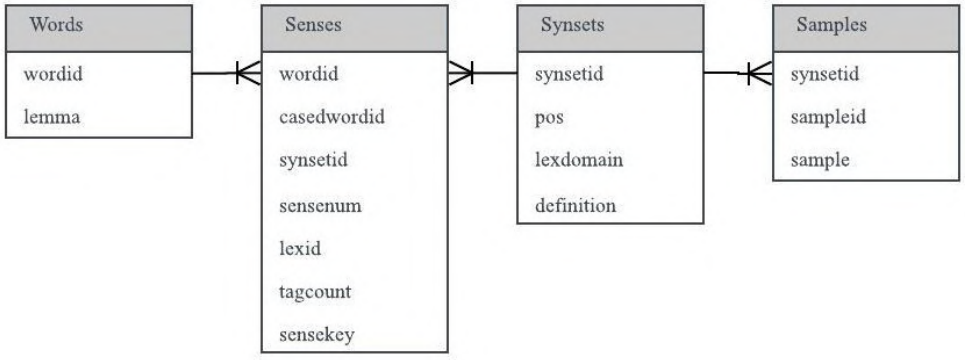

Figure 4 : ER Model of WordNet database used in Flying Dictionary

The database consists of 4 tables and they are as follows:

Words. This table consists of all the words present in the database. It has a list of all unique words characterized by a unique primary id for linking each word with other tables.

Senses. This table helps to create a link among other tables.

Synsets. This is the main table of the database which yields the definition and the classifications of the word (whether it is a verb, adjective, noun etc.). It also yields various synonyms of the words.

Samples. This table is a reference text corresponding to the meaning of the given word.

This way we create a steady and easy to access database. The SQL queries are written with concepts of joins to fetch data after linking the tables.

6. Evaluation

In order to evaluate the impact of Flying Dictionary on users' experience we conducted an online survey. The questionnaire was designed to elicit response from users on different aspects of their app -usage and new words learning experience during their course of usage. The questionnaire and user responses are reproduced in Table 1. It includes 7 psychometric questions based on 5- point Likert scales, and one preference-based question seeking the responder's overall preferred application on various criteria. The questionnaire was distributed to the users during the month of December. Users from all over the world participated in the feedback exercise to give a total of 59 useful responses that were consolidated. Table 1 below shows the questionnaire prepared for the purpose of our survey and the statistics derived from the responses garnered. Figure 5 shows their corresponding box and whisker plots.

1	2	3	4	5	Parameters
Q1) To what extent do you think the meanings of the words are accurate?					Median: 4 Mode: 4 InterQuartile Range: 1
Completely inaccurate: 10.2%	Somewhat inaccurate: 10.2%	Neither inaccurate nor accurate: 16.9%	Quite accurate: 33.9%	Completely accurate: 28.8%	
Q2) To what extent do you think the numbers of synonyms of the words are sufficient?					Median: 4 Mode: 4 InterQuartile Range: 1
Completely insufficient: 11.9%	Somewhat insufficient: 13.6%	Neither insufficient nor sufficient: 22%	Quite sufficient: 25.4%	Completely sufficient: 27.1%	
Q3) To what extent do you think the example sentences of the words are useful?					Median: 4 Mode: 3 InterQuartile Range: 1
Not at all useful: 15.3%	Less useful: 10.2%	Moderately useful: 18.6%	Somewhat useful: 28.8%	Very useful: 27.1%	
Q4) To what extent do you agree with the statement: "Flying Dictionary is user-friendly".					Median: 4 Mode: 4 InterQuartile Range: 2
Strongly disagree: 15.3%	Disagree: 8.5%	Neither disagree nor agree: 18.6%	Agree: 37.3%	Strongly agree: 20.3%	
Q5) How fast do you think is the search of words?					Median: 3 Mode: 4 InterQuartile Range: 2
Very slow: 13.6%	Slow: 11.9%	Neither slow nor fast: 20.3%	Fast: 38.8%	Very fast: 25.4%	
Q6) To what extent do you agree with the statement: "Flying Dictionary boosts my daily learning experience"					
Strongly disagree: 15.3%	Disagree: 10.2%	Neither disagree nor agree: 16.9%	Agree: 32.2%	Strongly agree: 25.4%	Median: 3 Mode: 3 InterQuartile Range: 2

<table>
<tr><td colspan="5">Q7) How frequently do you use the Flying Dictionary on a daily basis?</td><td rowspan="2">Median: 4
Mode: 4
InterQuartile Range: 2</td></tr>
<tr><td>Never:
15.3%</td><td>Rarely:
13.6%</td><td>Sometimes:
15.3%</td><td>Often:
33.9%</td><td>Regularly:
22%</td></tr>
<tr><td colspan="6">Q8) Choose your most preferred app among the following on the basis of:</td></tr>
<tr><td colspan="5">a) user-interface:
i. Flying Dictionary:
ii. onTouch Dictionary:
iii. Quick Dictionary:
iv. Google Chrome:
b) pop-up coming from as many apps as possible ? (e.g. PDFs, browsers, messaging apps, news apps etc.)
i. Flying Dictionary:
ii. onTouch Dictionary:
iii. Quick Dictionary:
iv. Google Chrome:
c) number of features being offered (e.g. (e.g. history, favorites, share, pronunciation, look-up on web, synonyms, example sentences etc.)
i. Flying Dictionary:
ii. onTouch Dictionary:
iii. Quick Dictionary:
iv. Google Chrome:
d) which app do you prefer the most overall?
i. Flying Dictionary:
ii. onTouch Dictionary:
iii. Quick Dictionary:
iv. Google Chrome:</td><td>

30.5%
16.9%
13.6%
39.0%

44.1%
13.6%
16.9%
25.4%

35.6%
23.7%
13.6%
27.1%

39.0%
15.3%
5.1%
40.7%</td></tr>
<tr><td colspan="5">Q9) On a scale of 1-5, with 1 being “poor” and 5 being “excellent,” how would you rate the ease of accessing the meanings of a word
a) in an app without the pop-up feature</td><td rowspan="2">Median: 3
Mode: 3
InterQuartile Range: 2</td></tr>
<tr><td>1:
10.1%</td><td>2:
20.3%</td><td>3:
42.4%</td><td>4:
25.4%</td><td>5:
1.7%</td></tr>
<tr><td colspan="5">b) in an app with the pop-up feature (Flying Dictionary)</td><td rowspan="2">Median: 4
Mode: 4
InterQuartile Range: 1</td></tr>
<tr><td>1:
1.7%</td><td>2:
0%</td><td>3:
23.7%</td><td>4:
50.8%</td><td>5:
25.4%</td></tr>
</table>

Q10) On a scale of 1-5, with 1 being "poor" and 5 being "excellent," how would you rate the availability of features like pronunciation and web look up a) in an app without the pop-up feature					Median: 3 Mode: 3 InterQuartile Range: 1
1: 1.7%	2: 13.5%	3: 38.9%	4: 37.2%	5: 8.5%	
b) in an app with the pop-up feature (Flying Dictionary)					Median: 4 Mode: 4 InterQuartile Range: 2
1: 1.7%	2: 6.8%	3: 20.3%	4: 42.3%	5: 28.8%	
Q11) On a scale of 1-5, with 1 being "poor" and 5 being "excellent," how would you rate the impact on battery drainage a) in an app without the pop-up feature					Median: 4 Mode: 4 InterQuartile Range: 2
1: 1.7%	2: 3.4%	3: 20.3%	4: 38.9%	5: 35.6%	
b) in an app with the pop-up feature (Flying Dictionary)					Median: 4 Mode: 4 InterQuartile Range: 2
1: 1.7%	2: 10.2%	3: 27.11%	4: 32.2%	5: 28.8%	

Table 1: Questionnaire on impact of Flying Dictionary and Response

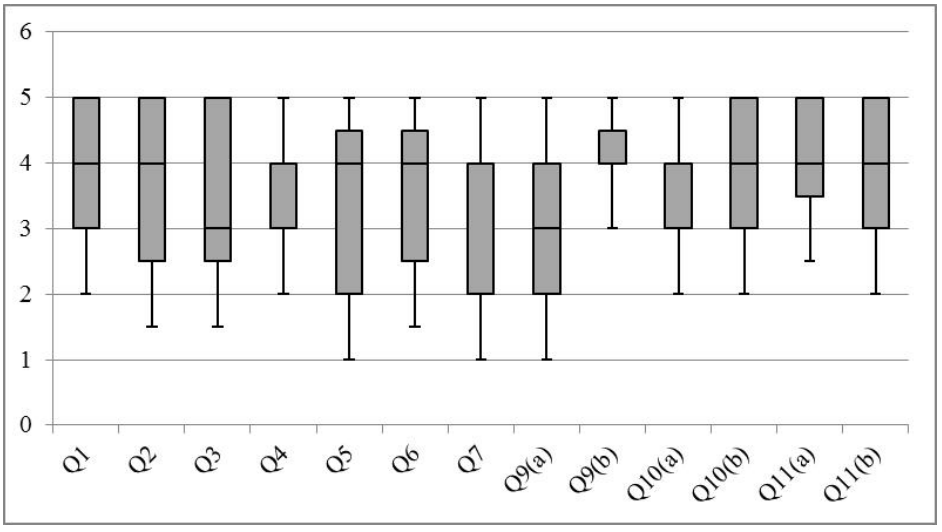

Figure 5: Box and whisker plots of the questions

In response to Q1, two-thirds of the users agree that he meanings of the words they found in the dictionary are accurate enough for their use. This sets a positive tone for evaluating the dictionary on various other parameters. The response to Q2 shows that around half the users felt that the synonyms were also at par, and as shown in Q3, about the same number of users felt that the example sentences that came in the dictionary were useful. However, about 20% of the users in each of these questions remained moderate in their answers while a few users also believed that these 3 parameters, i.e. meanings, synonyms, and examples were not sufficient enough. What emerges from these three answers is that the users are mainly content with the data and knowledge they are receiving from the application.

Q4 probes an important overall aspect of the overall user-experience of the app being user-friendly. 37.3% of the users agree with the app being friendly, while a whopping 20.3% strongly agree that the app is user-friendly. The rest do not agree with the statement though, and for them their needs will be catered to in the future updates. Another important aspect which comes with any android application and which also sets important criteria of enhancing the user -experience is the speed. In our case, the speed with which the app is able to fetch the words from the database is extremely important in order to boost our main aim of helping the lazy users' access meanings of words at the click of a button. The same issue was raised in Q5, which shows that about two-thirds of the users believe that the meanings come out fast enough.

Q6 probes whether the user feels that the use of the application is helping them boost their daily learning experience. 25.4% strongly agree and 32.2% agree with the statement, while the rest of the users could not fully agree to the statement. Q7 further delves into the question in order to judge their criteria of usage, whether the users are frequent users of the app in order to give their views. It was found that three-fourth of the users were frequent users while a major 15.3% of the users never used the app! The response shows that even though many of the users had the app in the phone, mostly they used the app, but many were present who didn't use it and were fine with their phone memory lying wasted without any use!

Q8 was asked to see which application the users preferred the most among the top contenders of the category on various classifications. The first criterion was of user-interface, it was found that 39.0% of the users preferred Google Chrome, while 30.5% users preferred Flying Dictionary's user-interface. That is probably due to the fact that Google can make its changes in the operating system itself after collaborating with Android, and we being an external app lack this feature. The second criterion was over how many apps they could use the pop-up feature. Here, a majority of 44.1% users believed that Flying Dictionary provided them with most possibilities of choosing from various apps. The third criterion was the number of features being offered along with the application. Here also most of the users believed that Flying Dictionary was doing a better job in the domain, compared to its other alternatives. The last criterion was their preferences based on the overall features. Here again we faced tough competition with Google Chrome, but perform well above in other aspects where other major apps in the league fail.

we asked three more questions with two parts each in order to look into the impact of the pop-up feature. These questions were focused on comparing apps which do not provide the pop-up feature vs. Flying Dictionary app which does provide the pop-up feature, on the basis of the major attributes

that are affected. A paired-samples t-test was conducted for each of these three questions to compare user -experience in no pop -up and pop-up conditions. The data analyzed is populated in Table 2. Also, a Mann-whitney U Test was also performed on these questions, whose data is populated in Table 3.

Student T Test	**Q9 (a)**	**Q9 (b)**	**Q10 (a)**	**Q10 (b)**	**Q11 (a)**	**Q11 (b)**
Mean Score	2.881356	4.016949	3.372881	3.898305	4.033898	3.627119
Standard Deviation	0.966414	0.7069	0.888593	0.959434	0.927847	1.039545
Difference	1.13559322		0.525423729		-0.271186441	
p-value	1.02446E-15		0.002519853		0.106587965	

Table 2. Results of performing Student T Test on Q9, Q10 and Q11

Mann Whitney U Test	**Q9 (a)**	**Q9 (b)**	**Q10 (a)**	**Q10 (b)**	**Q11 (a)**	**Q11 (b)**
Sum of Ranks	2431.5	4589.5	2949	4072	3766.5	3254.5
Mean of Ranks	41.21	77.79	49.98	69.02	63.84	55.16
Expected Sum of Ranks	3510.5	3510.5	3510.5	3510.5	3510.5	3510.5
Expected Mean of Ranks	59.5	59.5	59.5	59.5	59.5	59.5
U-value	2819.5	661.5	2302	1179	1484.5	1996.5
Expected U-value	1740.5	1740.5	1740.5	1740.5	1740.5	1740.5
U-value	661.5		1179		1484.5	
Z-score	-5.80477		-3.01945		1.37517	
p-value	<0.00001		0.00252		0.16758	

Table 3:Results of performing Mann-Whitney U Test on Q9, Q10 and Q11

Q9 dealt with the ease with which the meanings of the words were accessed in the two types of apps. while a majority 42.4% of the users rated 3 for apps without the pop-up feature, an even greater majority 50.8% of the users rated 4 for apps with the pop-up feature. Performing the paired-samples t-test revealed that there was a significant difference in the scores for apps without pop-up feature (M=2.881356, SD=0.966414) and apps without pop-up feature (M=4.016949, SD=0.7069); p=1.02446E -15. Also, the Mann-whitney U Test reveals a p -value of <0.00001 and since it is lesser than 0.05, hence this test also are in favor of the results being significant. These results suggest that the users definitely found that the pop-up feature eased the manner in which they looked for meanings. The p-value indicates that the results are highly significant.

Not only meanings, there are many occasions that the user may want to know the pronunciation of a word, or search the web for more information on it. In a typical situation, the user will have to switch to another app like a web browser. On the other hand, Flying Dictionary provides these features in the pop-up itself. Q10 deals with this functionality itself, asking the users to rate their experience with these features. The ratings received were mixed in the case of apps without the pop-up feature, whereas about half of the users scored a 4 or 5 in the case of Flying Dictionary. Performing the paired-samples t-test revealed that there was a significant difference in the scores for apps without pop-up feature (M=3.372881, SD=0.888592554) and apps without pop-up feature (M=3.898305, SD=0.959434375); p=0.002519853. Also, the Mann-whitney U Test reveals a p-value of 0.00252 which is lesser than 0.05. A low p-value in both the tests suggests that the result that features like pronunciation and web look-up are more readily available in Flying Dictionary is significant and not a null hypothesis.

Now since the pop-up feature requires a service to run in the background continuously to detect when a word has been copied, this can affect the battery usage in various phones according to their internal architecture. Q11 looks for such changes in battery performances. The mean score shifted from 4.033898 in apps without pop -up dictionary to 3.627119 in Flying Dictionary, showing a slight decrease in battery performance. But, a p-value of 0.106587965 in student t-test, and a p-value of 0.16758 in Mann-whitney U-test which is greater than 0.1 and 0.05, respectively, suggests that the results are not statistically significant and the impact on battery drainage cannot be accurately confirmed from these responses collected.

7. Conclusion

For the past two decades it has been recognized that the practice of using technology for learning purposes has seen a veritable explosion(w egner, 2003). The use of technology has not only created new opportunities within the traditional classroom, but has also served to expand learning experiences beyond the popular notion of the physical classroom with its traditional learning methods and tools, leading to an interesting, attractive and interactive media of learning and teaching (Serwatka, 2003). w ith developments involving a much more extensive e-learning approach, the use of smartphone-mediated dictionaries becomes

much more useful in such learning management systems. The results of this study confirmed previous results demonstrating advantages of pop-up dictionaries compared with dictionaries without the pop-up feature. The pop-up dictionary was found to be better due to the minimized extraneous cognitive load.

To conclude, the pop-up feature in Flying Dictionary brings significant improvement in the overall learning experience, reading comprehension and vocabulary development of the user. It not just provides them the meanings that they seek, but also provides them with ease of search so that they do not have to face any problems by switching their apps or in fact even going online to look up for the meanings. The pop-up feature indeed proves to be a better alternative in terms of enhancing the overall learning experience compared to the regular apps without the pop-up feature. The users of the application are satisfied with the functioning of the app, and it can be deduced that the application is ready to be used as a basis of learning at all platforms. It can be used by students of various institutes like schools and colleges, and can also be used by professionals, be it in their official work or be it in their daily use, the app proves to be an all-rounder.

In comparison to other dictionary apps already available, where one app might have one good feature and the other app some other good feature, Flying Dictionary tries to bring out all the best features together clubbed into one. Flying Dictionary provides all the features needed by the user to make their understanding and learning a seamless process by providing numerous features as already have been mentioned. From the experiment conducted it is deduced that among all the popular dictionaries with the pop-up feature, the users prefer Flying Dictionary; the only other tough competition being Google Chrome which comes handy while reading something from Google Chrome itself, and it also needs an online connection. For other textual content like PDFs and news apps, the users cannot use Google Chrome and hence they turn to using Flying Dictionary.

The need of the hour is to raise awareness about pop-up dictionaries among Android users, and incorporate pop-up feature in applications that are widely used because of the brand-names, like Oxford Dictionary and Merriam w ebster Dictionary, but which lack the pop-up feature, hence undermining the abilities of the apps which could have been phenomenal had they been incorporated with the pop-up feature. An even better initiative that can be taken is to incorporate the pop-up feature in the Android OS itself, so that users get it via the system itself, and no downloading of an external application is needed.

References

AbuSeileek, A. (2011). Hypermedia annotation presentation: The effect of location and type on the efl learners' achievement in reading comprehension and vocabulary acquisition, Computers & Education. 57(1):1281-1291. doi: 10.1016/j.compedu.2011.01.011

Alharbi, M. A. (2016). Using different types of dictionaries for improving EFL reading comprehension and vocabulary learning, The JALT Call Journal. 12(2):123-149. ISSN: 1832-4215

Anderson, J.R. (1995). Cognitive psychology and its implications (4th ed.). New York: Freeman

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory, Journal of Experimental Psychology: Learning, Memory, and Cognition. Vol. 33, pp. 570–585. doi: 10.1037/0278-7393.33.3.570

Chinetha, K. Joann, J. D. Shalini, A. (2015, February). An Evolution of Android Operating System and Its Version, International Journal of Engineering and Applied Sciences (IJEAS). 2(2):30-33. ISSN: 2394-3661

DictData. (2014). English Dictionary w ordNet (Version 1.5.4). [Mobile application software] Retrieved from https://play.google.com/store/apps/details?id=colordict.dictdata.dictionary.english.wordnet3

Farkade, A. M. Kaware, A. R. (2015, January). The Android - A w idely Growing Mobile Operating System w ith its Mobile based Applications, International Journal of Computer Science and Mobile Applications. 3(1):39-45. ISSN: 2321-8363

Filipwicz, L. (2016, October 7). Look Up replaces Define in iOS 10: Here's how to use it Retrieved December 31, 2017 from https://www.imore.com/look-replaces-define-ios-10-heres-how-use-it

Lin, H. F. (2007). The role of online and offline features in sustaining virtual communities: an empirical study", Internet Research, 17(2):119-138. doi: 10.1108/10662240710736997

Gordon, w . (2015, April 6). Chrome for Android Can Instantly Search Any Text You Highlight on Google Retrieved December 31, 2017 from https://lifehacker.com/chrome-for-android-can-instantly-search-any-text-you-hi-1709132529

Hulstijn, J.H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), Cognition and second language instruction. Cambridge University Press. pp. 258–286

Liu, T. & Lin, P. (2011). what comes with technological convenience? Exploring the behaviours and performances of learning with computer-mediated dictionaries, Computers in Human Behaviour. 27(1):373-383.

Manivannan, M. (2015, September 1). How many new English words are added to the dictionary every year? Retrieved December 25, 2017 from https://www.quora.com/How-many-new-English-words-are-added-to-the-dictionary-every-yearTuthor

T., Author A. and Buthor B. (1998a). An article in conference proceedings. In Maybe M., Some S. and Editor E., editors, Proc. of IDKw'98 (5 th conference), pp. 24-33.

Ozaz. (2014, December 14). Is there an app that can lookup word definition in

Android? Retrieved December 31, 2017 from https://forums.androidcentral.com/general-help-how/459510-there-app-can-lookup-word-definition-android.html

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. Instructional Science. Vol. 32, pp. 1–8

Patange, T. (2016, August 30). How Android Has Eaten Up Other Smartphone OSs: Is The iOS Next? Retrieved January 4, 2018 from https://dazeinfo.com/2016/08/30/android-smartphone-os-apple-ios-market-share/

Redkar, H. Bhingardive, S. Kanjojia, D. Bhattacharyya, P. (2015, January 25-30). w orld WordNet database structure: an efficient schema for storing information of WordNets of the world, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin, Texas

Rajarajeswari, M. Mohana, A. (2013, March). English Language: Growth of Vocabulary, International Journal on Studies in English Language and Literature (IJSELL). 1(1):41-47.

Schmitt, N. (2000). Vocabulary in language teaching. Cambridge: Cambridge University Press. ISBN: 0 521 660483

Serwatka, J. A. (2003). Internet-based Instruction in CIS Courses. Internet paper available at http://www.ihets.org/index.html

Shaunacy, F. (2017, April 17). The Most Common Second Languages Spoken Around the w orld Retrieved January 5, 2018 from http://mentalfloss.com/article/94456/most-common-second-languages-spoken-around-world

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load, Educational Psychology Review. Vol. 22, pp. 123-138. ISSN: ISSN-1040-726X.

Thamizharasi, R. (2016, February). Android Mobile Application Build on Android studio, International Journal of Modern Computer Science (IJMCS). 4(1):1-4

w egner, S.B., Holloway, K.C., Garton, E.M. (1999). The effects of Internet-based instruction on student learning, Journal of Asynchronous Learning Networks. Vol. 3, pp. 98–106.

Zhiliang, S. (2008). A comparative study of three learning strategies in efl students' incidental vocabulary acquisition, CELEA Journal. 31(6):91-101.

TICAM'18, p. 47 - 55

Dictionnaire des verbes d'informatique pour la traduction automatique en tamazight

Farida YAMOUNI
Université Mouloud Mammeri Tizi Ouzou Algérie
fariyamo@yahoo.fr

Résumé

Les logiciels de traduction automatique nécessitent de plus en plus de ressources terminologiques. Les travaux restent à faire pour les langues techniques et de spécialité. Dans cet article, nous présentons la description du dictionnaire des verbes d'informatique en Tamazight (kabyle) en vue de la traduction automatique. Nous élaborons dans NooJ[1] les ressources linguistiques utilisées pour l'analyse et la traduction puis nous décrivons les verbes recensés. Nous présentons ensuite les ressources mises en place. Dans Nooj, l'analyse linguistique est tout d'abord effectuée avant l'analyse syntaxique. Les dictionnaires électroniques existants (Aoughlis *et al.,* 2013) peuvent être utilisés avec celui des verbes construit en vue de la traduction automatique de terminologie informatique. NooJ est une plate-forme de développement permettant la description des langues naturelles ou techniques. Nos travaux entrent dans le cadre du système Nooj.

Mots-clés : Traduction automatique, NooJ, Tamazight.

1. Introduction

Des travaux récents ont été effectués sur les mots composés d'informatique en Tamazight (Yamouni, 2016) avec la construction dictionnaire des termes et la traduction automatique de termes dans NooJ.

Afin de construire le dictionnaire des verbes d'informatique en tamazight, un recensement des verbes est réalisé avec l'extraction manuelle des termes dans le lexique AMAw AL (Saad-Bouzefrane, 1996).

Nous nous intéressons dans cet article à la codification complète des verbes dans NooJ avec une traduction en français et en anglais. Nous pouvons aussi envisager la construction d'un dictionnaire avec les entrées en français ou anglais avec la traduction vers les deux autres langues.

Chaque entrée contient des informations sur la langue d'où a été créé ou extrait le verbe. Des tests sont réalisés pour la traduction automatique. Les résultats sont présentés pour quelques exemples.

1 http://www.nooj4nlp.net/

2. Le lexique AMAWAL

Il contient les termes d'informatique en tamazight. La décision de l'auteure (Saad-Bouzefrane, 1996) d'entreprendre la création d'un lexique d'informatique dans la langue amazighe a été motivée par le désir de contribuer à la connaissance et l'utilisation de la terminologie informatique en langue amazighe.

Les termes français et anglais du vocabulaire informatique ont été recensés par l'auteure, puis divers ouvrages tels que dictionnaires dans les différents parlers (kabyle, touareg, chleuh, mozabite, chaoui), lexiques Amawal, de mathématiques, d'électricité ont été utilisés afin d'avoir plus de possibilités de trouver un mot berbère à utiliser pour la traduction et n'envisager la néologie qu'en dernier recours.

Les termes d'informatique ont été associés à des traductions puisées dans les dictionnaires berbères confectionnés pour la majorité par des ethnologues. La majeure partie des mots berbères sont suivis d'une référence bibliographique mise en indice et du sens général du mot.

Par exemple, dans le lexique, Acheminer (to forward) : Seḍfeṛ (faire suivre KBL) KBL fait référence en bibliographie à (Dallet, 1982).

Parfois, bien que le terme berbère correspondant existe, l'auteure a utilisé un *autre terme plus adéquat. On citera le mot hachage qui devrait se traduire* par le mot ageddeḥ KBL(de geddeḥ: hâcher KBL), mais dans le contexte informatique, il a plus le sens de dispersion, d'où sa traduction par adway KBL (de dwi: disperser KBL).

Dans la figure 1, nous donnons un extrait du lexique d'informatique conçu par l'auteure.

3. Le Dictionnaire des verbes d'informatique en Tamazight

Les entrées du dictionnaire tamazight-Français-Anglais

Le dictionnaire contient environ 205 entrées. Chaque verbe est extrait du lexique AMAwAL (Saad-Bouzefrane, 1996).

Nous avons codifié chaque entrée dans le format Nooj, en rajoutant une traduction vers le français et l'anglais dans le but de la traduction automatique des verbes. La langue source est tamazight, les langues cibles possibles sont le français et l'anglais.

Dans un premier temps nous retiendrons tamazight comme langue source et le français comme langue cible pour la traduction automatique.

A partir de ce dictionnaire nous pouvons construire un autre dictionnaire avec pour langue source le français ou l'anglais.

- ***Abandon*** (*abort*): ***Tuggin*** (de *egg*: abandonner KBL) || ***Annuf*** (de *anef* KBL: laisser, abandonner)
- ***Abaque*** (*abacus*): ***Tafelwit*** (= tableau MAT, MW, pl. tifelwa)
- ***Abonné*** (*subscriber*): ***Ameltaɣ*** (pl. imeltaɣen, de *lteɣ* MW: adhérer)
- ***Abrégé*** (*abstract*): ***Agzul*** (=résumé MW, pl. igzulen, de *sewzel*: raccourcir KBL, *iwzil* KBL= *igzal* MCΓ = être court)
- ***Abréviation*** (*abbreviation*): ***Tazegzilt*** MW (pl. tizegzilin, v. abrégé)
- ***Accélérateur*** (*accelerator*): ***Ameckaḍ****** (pl. imeckaḍen, de *ckeḍ*: accélérer MAT)
- ***Accélérateur graphique*** (*graphics accelerator*): ***Ameckaḍ****** ***udlif****** (v. graphisme)
- ***Accès*** (*access*): ***Addaf*** (pl. addafen, de *atef* MZB,MCΓ = *adef* KBL = accéder, entrer), action d'accéder: ***adduf*** KBL, ***attaf*** MZB, MCΓ
  - ***Accès à distance*** (*remote access*): ***Addaf*** MZB ***agwemmaḍ*** KBL (de *agwemmeḍ*: l'autre côté de la valléee, versant de la montagne en face, éloigné) || ***Addaf anmeggag*** MCΓ (=éloigné, de *ugag*: être éloigné CLH,MCΓ)
  - ***Accès à l'Internet*** (*access to the Internet*): ***Addaf ɣer Internet***
  - ***Accès à une base de données distante*** (*remote database access*): ***Addaf ɣer taffa*** KBL ***tagwemmaṭ*** KBL ***n isefka*** (v. données) || ***Addaf ɣer taffa tanmeggagt*** MCΓ (éloignée, de *ugag*: être éloigné CLH,MCΓ) ***n isefka***
  - ***Accès au réseau à distance*** (*Remote Network Access, abr. RNA*): ***Addaf ɣer uzeṭṭa*** KBL ***agwemmaḍ*** KBL || ***Addaf ɣer uzeṭṭa anmeggag*** MCΓ (éloigné, de *ugag*: être éloigné CLH,MCΓ)
  - ***Accès complet*** (*full access*): ***Addaf ummid*** MAT (de *mmed* KBL, MAT: être complet)
  - ***Accès conflictuel*** [ou concurrent] (*concurrent access*): ***Addaf amgarrad*** (de *mgirred*: être en désaccord KBL)
  - ***Accès direct*** (*direct access*): ***Addaf usrid*** MW || ***Addaf anamad*** MCΓ (de *namad* MCΓ: se diriger vers, Rmq. *anamud* MCΓ: direction, *namad*: directement MCΓ)
  - ***Accès direct à la mémoire*** (*direct memory access*, abr. *DMA*): ***Addaf usrid*** MW ***ɣer tkatut*** MCΓ (v. mémoire) || ***Addaf anamad ɣer tkatut***
  - ***Accès en temps réel*** (*real-time demand*): ***Addaf s wakud*** MCΓ, MW ***ilaw*** MAT,MCΓ

Accès multivoie (*multichannel access*): ***Addaf agetbadu****** (de *aget* MAT: multi et *abadu* MCΓ: canal)

Figure 1 : Extrait du lexique d'informatique Français - Anglais - Berbère. (Saad-Bouzefrane, 1996)

3.2. Format d'une entrée

Dans l'extrait du dictionnaire des verbes d'informatique créé et présenté en figure 2,

pour l'entrée :

Beqqeḍ,V+MZB+FR=Afficher+EN=to display

- beqqeḍ est le verbe en tamazight,
- V indique la catégorie syntaxique de l'entrée,
- « +MZB » indique la référence bibliographique d'où a été extrait ou créé le terme, avec MZB: (Delheure, 1984)..
- « +FR= » permet de définir la traduction en français de l'entrée, ici afficher,
- « +EN= » pour donner la traduction en anglais de l'entrée, ici to display.

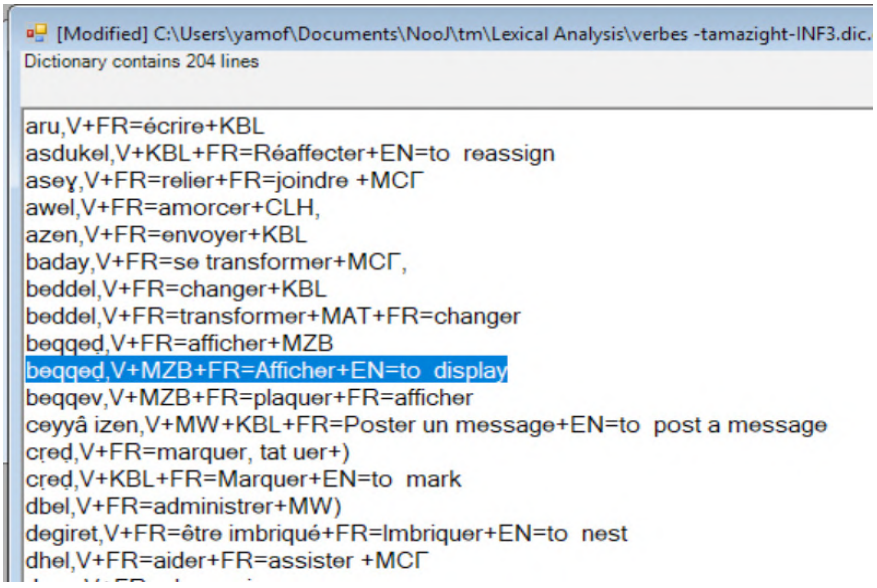

Figure 2 : Extrait du dictionnaire des verbes d'informatique : entrée beqqeḍ

3.3. Extrait du dictionnaire

Dans la figure 3, nous trouvons un extrait du dictionnaire des verbes d'informatique qui contient les termes recensés dans le lexique AMAwAL (Saad-Bouzefrane, 1996), puis codifiés dans NooJ. Actuellement, il y a 206 entrées.

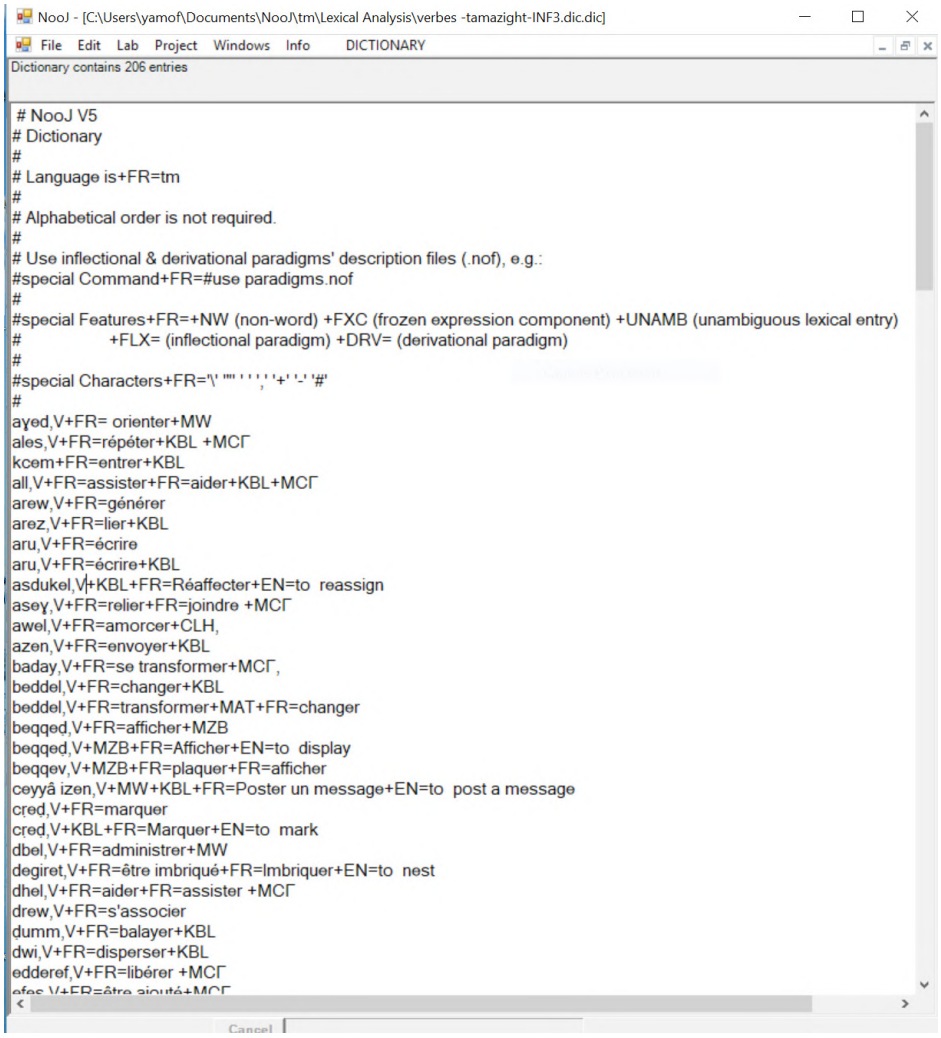

Figure 3 : Extrait du dictionnaire des verbes d'informatique.

4. La grammaire de traduction

Dans le système NooJ (Silberztein, 2003), les grammaires syntaxiques sont représentées par des ensembles structurés de graphes ou de règles qui décrivent des paradigmes syntaxiques stockés dans des fichiers «.nog «.

Pour l'analyse syntaxique automatique, nous avons construit une grammaire syntaxique pour la traduction automatique des verbes.

Une analyse linguistique est d'abord faite pour récupérer les catégories syntaxiques obtenues

grâce à la consultation des dictionnaires des deux langues.

Dans la figure 4, nous trouvons la grammaire de traduction automatique des verbes d'informatique de tamazight vers le français.

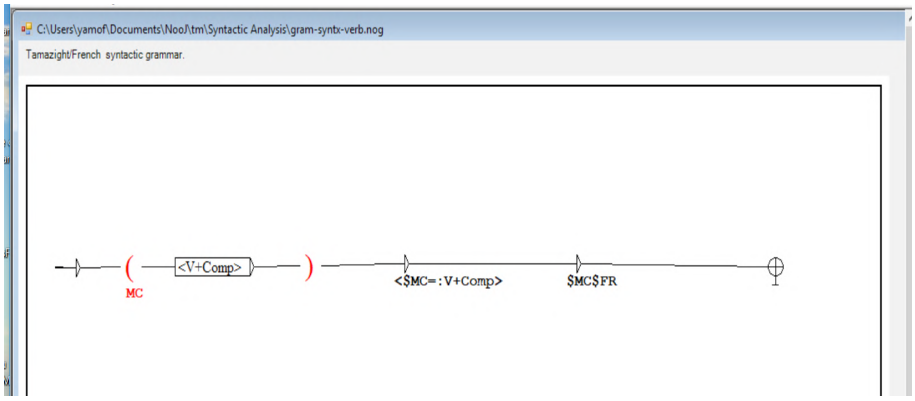

Figure 4 : La grammaire de traduction

5. Traduction automatique

5.1. La traduction automatique dans Nooj

Nooj est un environnement de développement linguistique qui permet de créer et utiliser des ressources linguistiques en vue du traitement automatique des langues naturelles et techniques.

Dans le cadre de la traduction automatique de terminologie informatique en tamazight, nous trouvons les travaux sur les mots composés d'informatique en tamazight (Yamouni, 2016). D'autres travaux ont par exemple été réalisés pour la langue arabe (Ferhi, 2012) pour la reconnaissance automatique des entités nommées arabes et leur traduction vers le français.

5.2. Un exemple de traduction automatique de verbe : cas de sermed

Nous donnons ci-dessous dans la figure 5, une exécution complète dans Nooj pour traduire le verbe « sermed ».

A partir du texte saisi dans le fichier « sermed.not », on lance l'analyse linguistique puis « locate pattern » en utilisant la grammaire Nooj de traduction « gram-syntx-verb.nog ». On affiche la concordance qui contient le verbe source et sa traduction, ici « sermed/activer » et « sermed/précipiter » car il y a deux traductions possibles codifiées dans le dictionnaire de traduction.

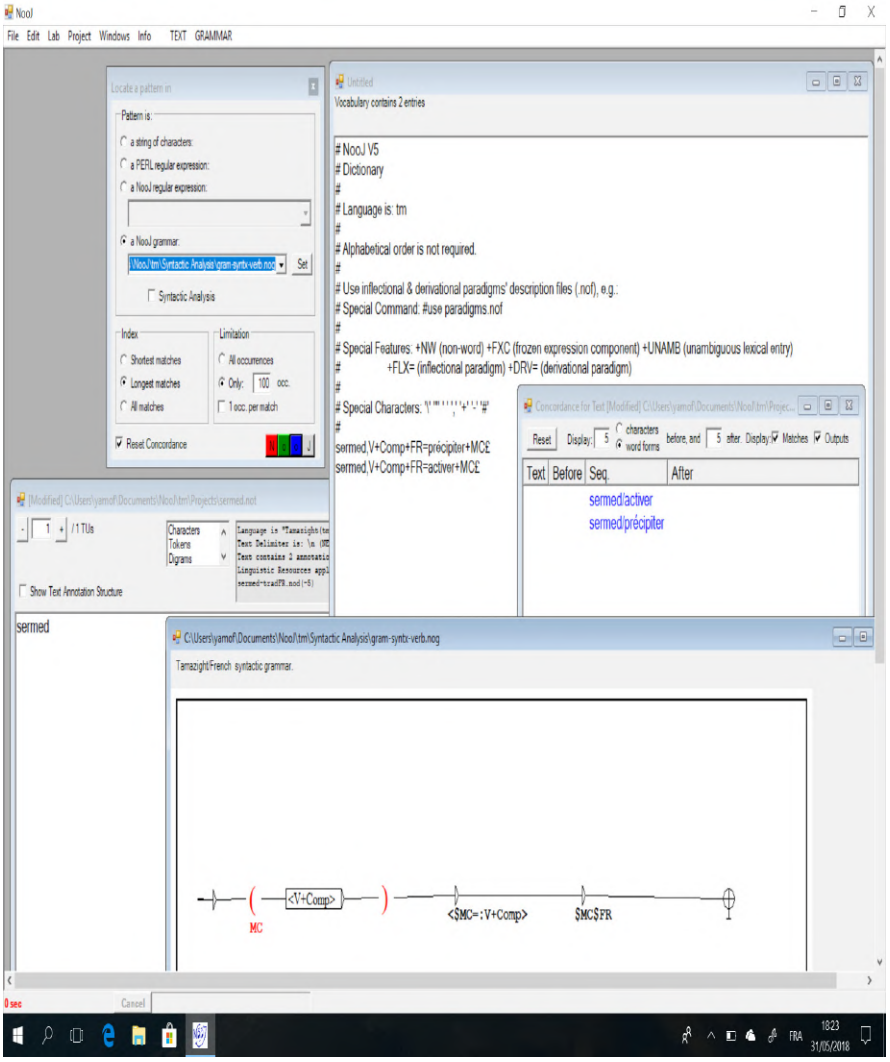

Figure 5 : Exemple de traduction automatique du verbe sermed

6. Conclusion et perspectives

Dans ce travail, nous avons étudié les verbes d'informatique en tamazight, nous avons construit les ressources linguistiques pour la traduction automatique des termes d'informatique de tamazight vers le français, en l'occurrence les verbes.

Les travaux réalisés permettent de traduire les verbes d'informatique. Les résultats obtenus permettent d'envisager un recensement d'autres verbes particulièrement ceux de l'internet, des systèmes embarqués, ….

Il est envisagé une étude détaillée de la conjugaison des verbes recensés avec la mise en place des règles de flexions dans le cas où celles élaborées dans (Aoughlis *et al.*, 2013) ne correspondent pas. Les résultats obtenus peuvent être utilisés par exemple pour concevoir et développer un outil pédagogique pour enseigner la terminologie informatique en tamazight.

Références

Aoughlis, F., Nait-Zerrad, K., Annouz, H., Aït-Kaci, F., Habet, M. S. (2013). A new Tamazight Module for NooJ. In Formalising Natural Languages with NooJ 2013. Selected Papers from the NooJ 2013 International Conference. Svetla Koeva, Slim Mesfar, Max Silberztein Eds. Cambridge Scholars Publishing: Newcastle upon Tyne.

Aoughlis, F. (2007). A Computer Science Electronic Dictionary for NOOJ. Lecture Notes in Computer Science 4592, Springer 2007, pp. 341-151.

Dallet, J. M. (1982). Dictionnaire Kabyle – Français : Parler des Ait Menguellat, SELAF, Paris.

Delheure, J. (1984). Dictionnaire Mozabite Français, SELAF, Paris.

Ferhi H. (2012). Reconnaissance automatique des entités nommées arabes et leur traduction vers le français. Thèse, 2012, Université de Sfax, Faculté des Sciences Economiques et de Gestion.

Silberztein, M. (2003). Manuel NOOJ. http://www.nooj4nlp.net

Saad-Bouzefrane S. (1996). Lexique d'informatique Français - Anglais – Berbère, (amawal n tsenselkimt Tafγansist - Taglizit – Tamaziγt), l'Harmattan Paris.

Yamouni F. (2016). A French-Tamazight MT System for Computer Science, 9th International Conference; NooJ 2015, Minsk; Belarus, June 11-13, 2015, revised selected papers. Automatic Processing of Natural-Language Electronic Texts with NooJ, Communications in Computer and information Science, Springer, Vol. 607, pp.208-217.

TICAM'18, p. 55 - 63

Base de données de toponymes d'Algérie : conception et réalisation

Noura TIGZIRI, Ramdane BOUKHERROUF
Laboratoire d'Aménagement et d'Enseignement de la langue Amazighes
Université Mouloud Mammeri de Tizi-Ouzou
{Nora.tigziri,ramdaneboukherrouf}@gmail.com

Résumé

Le travail présente quelques résultats préliminaires d'une base de données des toponymes de Kabylie pour la généraliser par la suite pour toute l'Algérie. Banque de données qui diffuse un certain nombre d'informations pour chaque nom de lieu intégré. Avec une normalisation d'écriture, le toponyme est identifié par une carte de géolocalisation suivie d'une fiche descriptive qui prend en charge son étymologie et sa signification, ses spécificités culturelles et historiques.

Mots-clés : Base de données, toponymie, Kabyle, carte de géolocalisation.

1. Introduction

La toponymie (du grec topo, lieu et onoma, nom) a pour objectif l'étude des noms de lieu ou toponymes, leur formation et leur évolution. C'est aussi l'ensemble des noms de lieux d'un pays, d'une région, d'une carte....

Les études toponymiques ont connu une grande évolution grâce à la collecte des données ou toponymes. Leur conservation dans des bases de données et leur représentation graphique sur des cartes numériques favorise grandement la conservation du patrimoine national.

Notre travail depuis quelques années en géographie linguistique (Tigziri, 2009 ; Tigziri et Boukherrouf, 2015 ; Tigziri et Boukherrouf, 2016) nous a confrontés à la problématique de la représentation d'un nom de lieu sur une carte. En effet, pour un lieu donné, il peut exister plus d'un nom connu pour le représenter et écrit avec différentes orthographes ce qui complique la recherche des paramètres de géolocalisation dans les bases de données de toponymes existant à l'heure actuelle. En effet, tant que l'on n'arrive pas à une normalisation de l'écriture des toponymes, ce qui est le cas dans notre pays, nous continuerons à nous débattre dans des problèmes de représentation spatiale de ces données.

Ce problème est tellement important qu'une première conférence sur la normalisation des noms géographiques a eu lieu à l'ONU, en 1967. Quelques années plus tard, un séminaire de toponymie franco-africain a été organisé en France, en juin 1996 à l'Unesco, dont l'objectif était d'inventorier les problèmes qui se posent aux pays africains dont l'étude des problèmes de normalisation et de transcription qui nous intéresse dans le cas de notre recherche.

Le Décret nº 81-27 du 7 mars 1981 portant établissement d'un lexique national des noms de villes, villages et autres lieux et cité plus haut «règlemente» d'une certaine manière la pratique et fait figure de politique toponymique du pays. L'article premier du texte décrète :

Article 1er

Les assemblées populaires communales sont chargées :

- *d'étudier et d'arrêter, de manière précise, la dénomination de tous les lieux possédant déjà un nom,*
- *de revoir certaines dénominations non conformes à nos traditions et de prévoir, le cas échéant, une nouvelle dénomination adaptée aux spécificités locales.*

Elles peuvent faire appel à toute personne, qui, en raison de sa compétence et/ou de son expérience est susceptible d'apporter un concours utile.

Force est de constater que malgré ces recommandations, on continue encore à retrouver des toponymes écrits sous différentes formes ce qui complique la tâche de la cartographie numérique d'un pays à l'aire de l'internet et du GPS.

Aussi nous proposons un projet intitulé Base de *données toponymiques de l'Algérie* qui peut cibler de nombreux objectifs :

- Normaliser l'écriture des toponymes en se basant sur des règles d'écriture élaborées par la commission de toponymie qui est l'organisme responsable de la gestion des noms de lieux du Québec. Elle a été créée en 1977, en vertu de l'article 122 de la Charte de la langue française. Elle prenait alors le relais de la Commission de géographie (1912-1977), des noms officiels (entités administratives, Communes, Daïras et Wilayas) et non officiels (lieux dits, hameaux...) et sur les spécificités de notre pays.
- Associer à chaque toponyme ses coordonnées de géolocalisation afin de faciliter sa représentation sur carte et l'accès aux données GPS.
- Associer à chaque toponyme, ses spécificités artisanales, culturelles, naturelles afin de faire connaître et revaloriser notre patrimoine culturel, artistique, artisanal ...et par conséquent touristique.

Ce travail sur la toponymie est aussi un moyen de réactiver des termes amazighes tombés en désuétude car comme nous le savons tous, les noms de lieux sont les éléments qui résistent le plus au changement. Par conséquent dans le cadre de l'aménagement de la langue amazighe, ces recherches pourront être d'une grande contribution.

2. Les toponymes algériens : quelques données linguistiques

Les toponymes nord-africains (Pellegrin, 1949) trouvent leurs origines dans le pré-berbère, le lybico-berbère, le phénicien, le grec et le latin, l'arabe ainsi qu'à moindre mesure l'espagnol et l'italien. Il ne faut pas oublier aussi l'apport du portugais, du turc et du français. Chaque occupant y a laissé sa trace. Mais les toponymes sont en majorité berbères et arabes. On retrouve les toponymes berbères presque partout et on les reconnait assez facilement grâce à leurs préfixes et leurs désinences et à la présence de t.....t, indice du féminin en berbère.

Les populations arabophones ont arabisé des toponymes pour créer parfois des toponymes-pléonasmes tels que comme Oued Souf (souf=asif), Bir Ghbalou (puits= aghbalou)), Ighil, Draâ...

Les français eux aussi ont apporté leur contribution dans le changement des toponymes. Ainsi à Alger Tala Umelil est devenu Télemly et Tagarins provient de tigrine (pente douce plantée de céréales).

Les Français comme les Arabes ont changé des toponymes mais en ont inventé d'autres. Parmi les plus récents, on citera Skikda qui est devenu Soukaikida.

Cet état de fait historique a engendré plusieurs amalgames et foisonnements relatifs à l'identité de la toponymie de l'Afrique du nord, en ce sens qu'elle touche à l'étymologie des toponymes, leur morphologie et leur transcription.

Les recherches berbérisantes ont suscité la curiosité de nombreux chercheurs pour tenter de traiter la problématique de la toponymie de l'Afrique du Nord sous plusieurs angles. Plusieurs travaux ont montré le rôle et l'apport de la toponymie dans les études linguistiques diachroniques du domaine amazighe (Cheriguen, 1993 ; Allati, 1998 ; Zakara, 1999 ; Ait Said, 2001).

La question de la transcription des toponymes a été évoquée depuis les années quarante par Basset (1942) qui a suggéré la nécessité d'adopter une forme unique pour un même toponyme géographique. Aussi, les travaux de Atoui (1998) et ceux d'Atoui et Benramdane (2005) ont montré clairement la problématique de la normalisation de l'écriture des toponymes algériens.

Dans la perspective de prendre en charge ces différentes contraintes soulignées par de nombreux travaux, Nehali (2013), nous tentons d'élaborer une base de données toponymiques, banque de données qui vise à recenser un maximum de toponymes avec leur géolocalisation.

Notre travail s'inscrit comme le prolongement de ces travaux en prenant en charge leur normalisation, leur géolocalisation exacte, leur signification et leurs caractéristiques historiques et culturelles.

3. La transcription des toponymes : quelques contraintes

Comme nous l'avons signalé dans le point précédent, la question de la transcription des toponymes en Algérie a montré une dispersion énorme pour chaque appellation. En effet, en plus de l'attribution de plusieurs appellations à chaque toponyme, la même appellation enregistre plusieurs variations morphologiques et de transcriptions. Ainsi par exemple, nous avons pour un même toponyme les écritures suivantes :

- Ait Iraten, At Iraten, Ait-Iraten, At-Iraten, At Irathen, AitIrathen, At-Irathen…
- Ait Mellal, At Mellal, Ait-Mellal, At-Mellal, Ait Mellel, AtMelel…

Cette dispersion pose des problèmes de géolocalisation et d'identification de la forme exacte du toponyme. C'est pourquoi avant de passer à la présentation de l'aspect technique de la base de données, il est important de présenter au préalable les choix adoptés en matière de la normalisation de l'écriture des toponymes amazighes. Pour ce faire, nous ferons appel aux

recommandations élaborées par la commission de toponymie qui est l'organisme responsable de la gestion des noms de lieux du Québec.

4. Conception de la base de données

Pour nous permettre de prendre en charge les différents champs évoqués dans l'introduction, la base de données va se présenter sous la forme de l'organigramme ci-dessous (figure 01) : Pour chaque toponyme, on lui associe toutes les spécificités culturelles, historiques, artisanales, etc. Chaque toponyme aura une écriture normalisée, son étymologie à cause d'une éventuelle transformation de son écriture au fil du temps. Chaque terme ainsi reconstitué peut-être réactivé s'il n'existe plus et le réinjecter dans les dictionnaires de la langue amazighe.

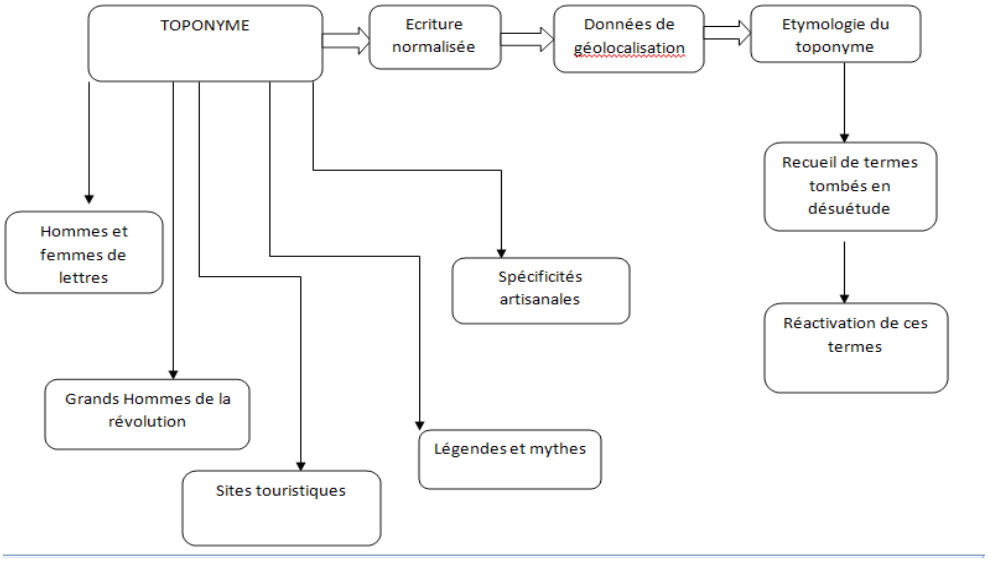

Figure 1 : Organigramme de la base de données toponymique

4.1. La géolocalisation des données toponymiques

La représentation spatiale de ces toponymes et de la variation phonétique et morphologique, la définition des coordonnées de ces points (longitude et latitude) n'est pas une tâche facile. En effet, les toponymes présentent une grande variation dans le temps et dans l'espace. Il nous arrive de ne pas pouvoir situer exactement un point d'enquête sur la carte parce que le nom a changé ou a été transformé. En effet, les diverses sources (cartes topographiques, enquêtes de Basset, documents administratifs fournis par la w ilaya) présentent parfois, des variations importantes dans l'écriture des toponymes et ceci est une difficulté supplémentaire à surmonter quand on passe à une représentation cartographique.

Par ailleurs, pour les points qui ne figurent pas sur la base de données toponymique de l'Algérie, nous avons fait appel à *Google Earth* pour extraire les données de géolocalisation. Nous donnons ci-dessous des exemples toponymes avec les coordonnées de localisation (Tableau 1).

Latitude	Longitude	Name
36,819	*4,213*	Abizar
36,823	*4,292*	Adrar n At Qdiɛa
36,811	*4,172*	Afir
36,614	*4,042*	Aglagal
36,809	*4,322*	Aγrib

Tableau 1 : Toponymes avec des données de géolocalisation

4.2. Présentation cartographique

Une fois les données de géolocalisation (latitude et longitude) définies pour les points d'enquête ciblés, nous les représentons sur une carte à l'aide du logiciel QGIS[1] (Geographic Information System) (Figure 2).

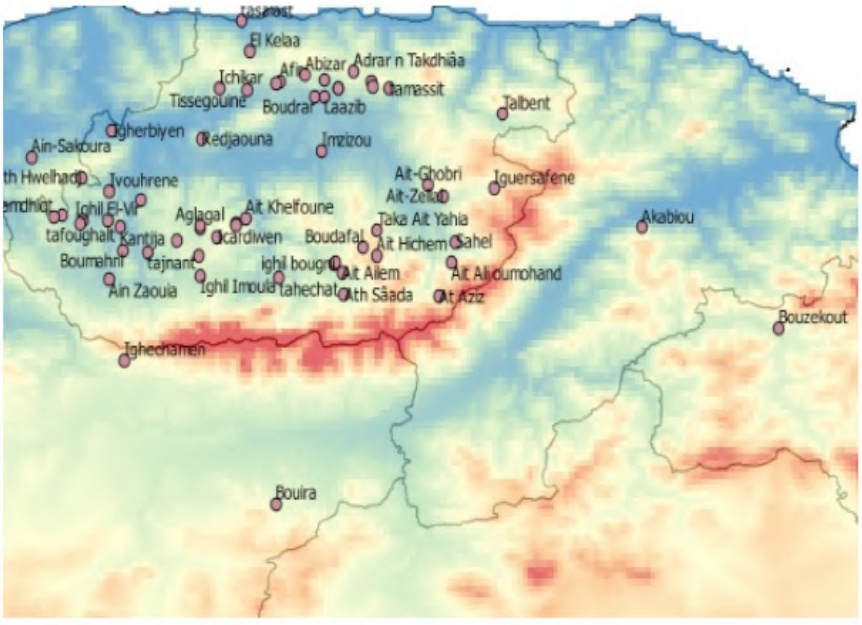

Figure 2 : Présentation cartographique des toponymes

1 Un logiciel de cartographie numérique accessible gratuitement sur le site : http://www.qgis.org/fr/site/

5. Présentation de la base de données

La base de données des toponymes est conçue avec comme objectif de diffuser un certain nombre d'informations pour chaque nom de lieu intégré. Avec une normalisation d'écriture, le toponyme est identifié par une carte de géolocalisation suivie d'une fiche descriptive qui prend en charge son étymologie et sa signification, ses spécificités culturelles et historiques.

La base de données présente la première interface qui permet de chercher dans la base de données le toponyme qu'on veut chercher (Figure 3) :

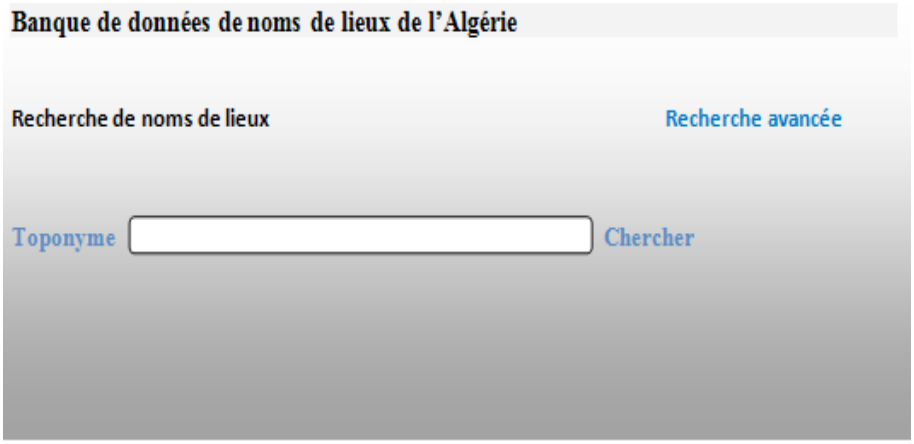

Figure 3 : Interface de la base de données

Si on veut faire une recherche accompagnée de quelques informations supplémentaires, il suffit de cliquer sur le bouton *recherche avancée*. Ainsi, l'interface sera affichée comme suit (Figure 4) :

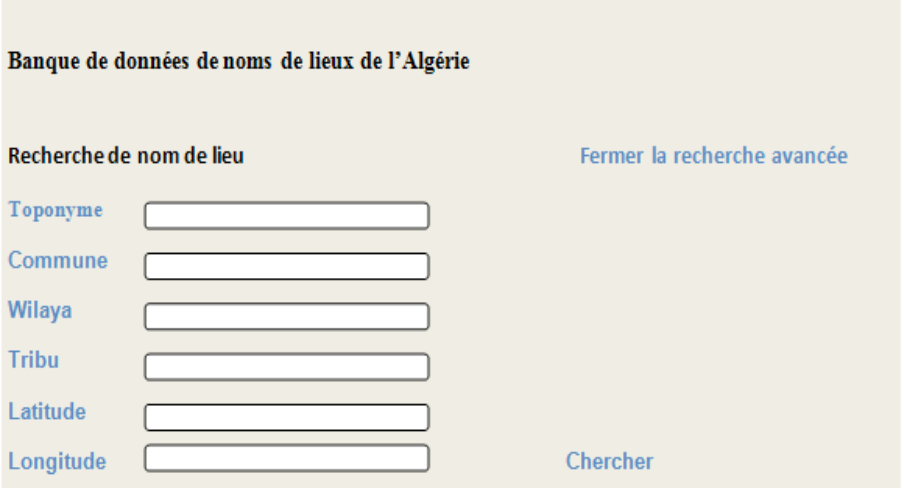

Figure 4 : Interface de la recherche avancée

Dans les deux cas de figure, la recherche d'un toponyme nous oriente vers la fiche descriptive (Figure 5), fiche qui prend en charge le nom de lieu, le type (village, commune, etc.), son appartenance administrative, sa tribu et ses caractéristiques historiques et culturelles.

Nom de lieux	Type	appartenance administrative	Tribu	coordonnées	caractéristiques

Figure 5 : Fiche descriptive des toponymes

Il suffit de cliquer sur le nom de lieu pour que la plate-forme nous affiche les détails de la fiche descriptive du toponyme, accompagnée de la carte de géolocalisation (figure 6).

Fiche descriptive du toponyme	
Carte de géolocalisation	
Nom de lieu	
Etymologie (signification)	
Type (Village, commune, wilaya, etc.)	
Tribu	
Coordonnées de géolocalisation	
Caractéristiques culturelles et historiques	

Figure 6 : Fiche descriptive du toponyme

6. Conclusion

Comme synthèse de notre travail, qui est encore en phase d'élaboration, nous présentons quelques résultats préliminaires d'une base de données des toponymes de l'Algérie, banque de données qui propose une normalisation d'écriture du toponyme identifié par une carte de géolocalisation suivie d'une fiche descriptive qui prend en charge son étymologie et sa signification, ses spécificités culturelles et historiques. Cette recherche sera aussi d'un grand apport pour l'aménagement de l'amazighe avec la réactivation de termes véhiculés par les toponymes mais tombés en désuétude.

Références

Aghali, M. Z. (1999). « Anthroponymes et Toponymes Touareg. Inventaire et corrélation », in *Littérature arabo-berbère. Dialectologie, ethnologie*, ERS 1723, CNRS 27-1999, Paris, pp. 209-248.

Ait Said, F. (2001). De *l'analyse des Toponymes berbères à travers des sources d'Al-Bakri (XIe siècle)*, Mémoire de DEA, INALCO, Paris.

Allati, A. (1998) « Tal : une base toponymique ancienne de l'Afrique du Nord et des îles Canaries », in NRO n° 31-32, Paris, pp. 143-156.

Atoui, B., Benramdane, F. (2005). « *Mondialisation et normalisation des toponymes et des écritures : le cas de l'Algérie* », in Nomination et dénomination. Des noms de lieux, de tribus et de personnes en Algérie, Coordonné par F. Benramdane et B. Atoui, édition du CRASC, Oran, 2005, pp. 187-197.

Atoui, B. (1998). *Toponymie et espace en Algérie, Institut National de Cartographie*, Alger.

Basset, A. (1942). *Note sur la graphie des toponymes (extrait des travaux de l'Institut de Recherche Sahariennes*, tome I, Alger, Imprimerie Imbert.

Boukherrouf, R., Tigziri, N. (2015). « Base de données kabyles : collectes de données et applications. Synchronisation texte / son », *Iles d imesli* N° 07, LAELA, UMMTO, pp. 193-206. http://revue.ummto.dz/index.php/idi/issue/view/122

Chaurand, J., Cheriguen, F. (1993), Toponymie algérienne des lieux habités (les noms composés), *In: Nouvelle revue d'onomastique*, n°23-24, pp. 259-260.

Jolivet, R. (2009). « Représentation spatiale des données d'enquête : outils informatiques pour une analyse exploratoire », *Revue Iles d Imesli*, N°01, Algérie, Université Mouloud Mammeri de Tizi-Ouzou, pp. 39-57. http://revue.ummto.dz/index.php/idi/article/view/205

Pellegrin, A. (1949). *Essai sur les noms de lieux d'Algérie et de Tunisie. Etymologie, signification*.

Tidjet, M., Nehali, D. (2013). « Initiation d'une base de données toponymique », Revue Iles d Imesli, N°05, Lalela-UMMTO, pp. 323-340.

Tigziri, N. (2009). « Carte phonétique du (l) », *Revue Iles d Imesli*, N°01, Algérie, Université Mouloud Mammeri de Tizi-Ouzou, pp. 29-37. http://revue.ummto.dz/index.php/idi/article/view/204

Tigziri, N. (2014). « La réalisation de grands corpus berbères normalisés et interopérables : enjeu culturel et enjeu d'ingénierie linguistique », *ASINAG N° 9*, Rabat, Maroc, IRCAM.

Tigziri, N. (2016). « Les corpus oraux : essai de segmentation automatique », *in Actes de la 3ème conférence internationale sur les technologies de l'information et de la communication et l'Amazighe*, Rabat, Maroc, IRCAM.

TICAM'18, p. 65 - 74

La traduction assistée par ordinateur : quelle utilité pour la langue amazighe ?

Fadoua ATAA ALLAH, Siham BOULAKNADEL
Centre des Etudes Informatiques, des Systèmes d'Information et de Communication
Institut Royal de la Culture Amazighe
{ataaallah,boulaknadel}@ircam.ma

Résumé

L'officialisation de la langue amazighe a fait naître des besoins grandissants en termes de traduction. En effet, depuis la constitutionnalisation de l'amazighe, en 2011, de nombreux organismes étatiques, privés et sociaux ont, progressivement, opté pour mettre à jour et localiser leur documentation. Face à cette demande grandissante, l'Institut Royal de la Culture Amazighe (IRCAM) s'est également démultiplié les efforts pour réussir cette phase de transition entre la reconnaissance et la généralisation de la langue à tous les niveaux. A cet effet, l'IRCAM a dédié un centre spécialisé en traduction pour rendre service au public, et assure des formations en traduction vers l'amazighe aux spécialistes. En outre, il s'est intéressé à la traduction par ordinateur pour améliorer ce service. Une telle démarche nécessite une étude sur la théorie de la traduction et sur les outils à déployer pour cet objectif. Dans ce contexte, cet article présente l'état actuel de la traduction par ordinateur de l'amazighe au Maroc, et ce selon trois niveaux : l'enseignement supérieur, le monde professionnel et de la recherche. Il introduit l'utilité de la traduction assistée par ordinateur pour la langue amazighe. Par ailleurs, l'article brosse un tableau d'outils de la traduction assistée par ordinateur et analyse leurs avantages et leurs inconvénients, afin de déterminer de quelle manière ils peuvent s'intégrer au mieux dans l'environnement du traducteur.

Mots clés : Langue Amazighe, TAO,THAO, mémoire de traduction.

1. Introduction

Le plurilinguisme, quoique qu'il représente une richesse linguistique et culturelle irréfutable, il revêt, à l'heure de la mondialisation, un défi et un enjeu majeurs pour les sociétés. En effet, la non-connaissance d'une langue peut engendrer un accès limité à l'information et constituer un frein au développement du potentiel intellectuel. Par conséquent, les communautés linguistiques avec un faible pouvoir économique souffrent souvent de cette discrimination.

Dans le contexte marocain, la gestion du plurilinguisme revêt aussi une dimension démocratique : il s'agit d'assurer à chaque citoyen l'accès aux services administratifs et aux textes législatifs dans sa langue première, afin qu'il ait connaissance de ses droits et puisse bénéficier des services de l'Etat dans une langue qu'il maîtrise. Certes, cette démarche a un coût de traduction et d'interprétation notable. Pour faire face à ce coût social et économique, des recherches ont été menées dans le but d'activer et d'avancer le processus de la traduction humaine.

Aujourd'hui, il existe toute une industrie consacrée à cette problématique. Elle fourni des prestations de traduction humaine mais aussi toute une gamme d'outils dans le but de diminuer les coûts de traduction. Dans ce cadre s'inscrit ce travail de recherche. Il vise de jeter la lumière sur la traduction par ordinateur au Maroc, et proposer des méthodologies à mettre en œuvre, notamment pour la langue amazighe.

Le reste de cet article est organisé comme suit. La section 2 introduit le domaine de la traduction par ordinateur. Elle présente un aperçu de la situation de la traduction par ordinateur de l'amazighe dans le contexte marocain. La section 3 dresse un ensemble de problèmes soulevés et de propositions envisagées. La section 4 expose le fonctionnement des outils de traduction humaine assistée par ordinateur (THAO). Finalement, la section 5 présente une étude comparative de neuf outils de THAO existants sur le marché.

2. Etat de la traduction par ordinateur de l'amazighe au Maroc

Dans la perspective d'examiner l'état de la traduction par ordinateur de l'amazighe au Maroc, il paraît primordial de distinguer entre les différentes catégories de traduction qui s'insèrent, selon le degré d'automatisation, dans un continuum linéaire (Hutchins et Somers, 1992) :

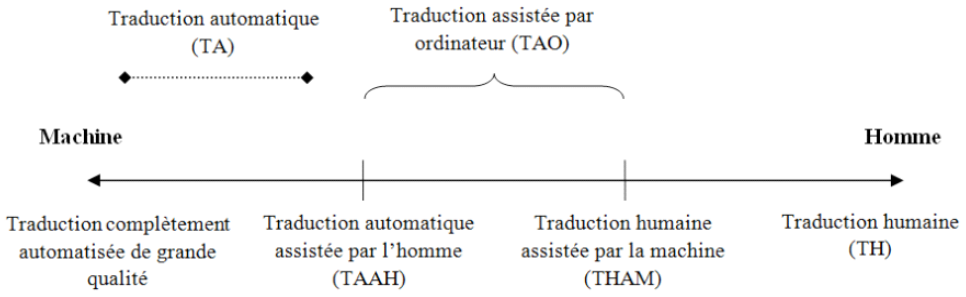

Figure 1 : Représentation des différents types de traduction[1]

- **Traduction humaine (TH)** : elle représente, tout simplement, la traduction faite par l'homme sans aucune intervention ou aide de la machine.
- **Traduction assistée par ordinateur (TAO)** : dans cette catégorie différents formes de collaboration entre l'homme et la machine interviennent. Cette collaboration peut être dirigée par l'homme ou par la machine, ce qui permet de distinguer entre deux types d'approches :
    - **Traduction humaine assistée par la machine (THAM)** : dans cette approche de traduction, le traducteur humain joue un rôle nettement plus grand. Il se sert d'un ou de plusieurs programmes informatiques pour l'assister dans une partie du processus de traduction.

1 La représentation proposée par Hutchins et Somers, extraite du (Quah, 2013).

- **Traduction automatique assistée par l'homme (TAAH)** : l'approche de la traduction assistée par l'homme est apparue pour désigner la traduction informatisée où le savoir humain est sollicité. L'intervention humaine peut être effectuée soit à la fin du processus comme une révision du produit finale ou bien au cours du traitement, autant de fois que ceci est nécessaire.

• **Traduction automatique (TA)** : dans cette catégorie, le processus de traduction consiste à utiliser un système informatique pour traduire un texte d'une langue naturelle vers une autre sans aucune intervention humaine.

2.1. Domaine de l'enseignement supérieur

Pour faire face aux nouveaux défis du 21e siècle, marqués par la révolution technologique et engendrés par le phénomène de la mondialisation, l'enseignement de la traduction à l'université, sur le plan international, s'est focalisé, en outre, sur une sous-discipline des études qui porte sur les technologies de la traduction. Cette dernière se donne pour objectif l'analyse des différentes contributions scientifiques concernant l'application, la création ou l'étude d'outils informatiques pour la traduction, afin de proposer une formation professionnalisante qui correspond aux besoins du marché (Arrouart, 2003 ; Pym, 2007).

Dans le contexte marocain, cette nouvelle tendance a également apparu dans l'enseignement supérieur où se révèle un intérêt croissant aux outils technologiques et à leurs applications. En effet, aux universités marocaines, l'enseignement de la traduction par ordinateur a été intégré au cursus, dont l'objectif est la nécessité d'emboiter le pas aux nouvelles technologies appliquées à la traduction et de contribuer ainsi à la formation d'un nouveau profil de traducteur. Malheureusement, à notre connaissance, l'enseignement de la traduction par ordinateur pour l'amazighe n'est pas encore intégré au cursus.

2.2. Domaine professionnel

Pour suivre les évolutions de la demande du marché de traduction qui augmente d'année en année, les agences de traduction, sur le plan international, exploitent plusieurs outils informatiques dans leur processus afin de bénéficier des avantages des uns et des autres. Néanmoins, les traducteurs indépendants trouvent que l'usage de ces outils n'est pas tellement rentable, dû au coût de l'acquisition et des mises à jour.

En ce qui concerne les agences marocaines consultées dans le cadre d'une enquête[2], il a été constaté que la majorité n'utilise pas les outils de traduction, et ce pour des raisons de limites budgétaires, notamment que le recours aux traducteurs indépendants est indispensable que ça soit pour la langue amazighe ou arabe.

2 Cette enquête a été menée sur le territoire marocain, auprès de 20 agences de traduction. Le questionnaire exploité dans cette enquête est explicité dans la section Annexe.

2.3. Domaine de la recherche

Le domaine de la traduction par ordinateur attire l'attention de nombreux chercheurs dans la sphère scientifique. Ces dernières décennies, la recherche en ce domaine a connu une évolution importante. D'ailleurs, les résultats, à l'échelle internationale, sont impressionnants, spécialement avec l'avènement de l'apprentissage profond.

A l'échelle nationale, les projets de recherche de grande envergure en traduction par ordinateur pour les langues d'usage, notamment l'arabe et l'amazighe, restent timides.

Concernant l'amazighe, l'IRCAM accorde une importance aux recherches menées dans le domaine de la traduction par ordinateur. Ces recherches attestent d'une tendance positive vers le développement de ressources et outils pour la traduction automatique pour l'amazighe (Miftah *et al.*, 2017; Taghbalout *et al.*, 2018).

3. Entraves et solution envisagée

3.1. Problèmes relevés

Après ce bref aperçu, il appert que dans le domaine de l'enseignement, notamment dans les universités marocaines, la plupart des filières n'exploitent pas les outils de traduction pour des raisons de limites budgétaires ou par manque d'expérience en ces outils. Quant aux agences de traduction, il est clair que le concept de l'usage des outils de traduction n'est pas encore bien implanté dans le milieu professionnel.

Dans le domaine de la recherche, des progrès ont été effectués au niveau de la traduction de la langue amazighe. Néanmoins, les ressources disponibles actuellement ne sont pas suffisantes pour une traduction automatique acceptable. Certes, il y a un intérêt remarqué et une évolution réalisée. Cependant, il faudrait avoir fort à faire pour promouvoir la recherche dans ce domaine au niveau des institutions aussi bien publiques que privées.

3.2. Solution envisagée

Face à ces entraves, il s'avère judicieux d'adapter la démarche du traducteur en exploitant les outils de la traduction assistée par ordinateur. Ces derniers assurent une rentabilité des données, en recyclant automatiquement les traductions antérieures ; une réduction du temps de réalisation, en disposant de traductions éventuelles ; et une cohérence des traductions, en harmonisant la terminologie.

Sans doute, bénéficiant des exemples de traduction, le traducteur aura l'avantage d'enrichir ses propos. Ainsi, disposant de plus de temps, il sera en mesure de prendre en charge des volumes beaucoup plus importants. En outre, profitant d'une terminologie harmonisée, la qualité des traductions sera améliorée non seulement par l'homogénéité des traductions antérieures mais aussi par l'homogénéité des traductions en équipe, notamment grâce à une centralisation des données. De ce fait, la traduction humaine assistée par ordinateur (THAO) permet à priori d'optimiser la productivité de ses utilisateurs et la rentabilité de leurs traductions.

4. Fonctionnement des outils de THAO

La traduction humaine assistée par ordinateur se base sur l'expertise humaine et se fixe pour objectif d'aider le traducteur dans sa mission à travers des outils. Ces derniers, connotés « outils à mémoire de traduction » ou « outils d'aide à la traduction », portent sur le concept du morcellement. Il consiste à segmenter le texte à traduire en éléments assez courts, afin d'assurer la présence d'éléments déjà traduits. A la base de ce concept, les outils d'aide à la traduction constituent des mémoires, qui permettent de réutiliser des phrases ou des segments de phrases traduits antérieurement.

D'ailleurs, les premiers systèmes, apparus en 1994, se base essentiellement sur les mémoires de traduction (Brace, 1994). Cependant de nos jours, les outils d'aide à la traduction sont enrichis par d'autres fonctionnalités, notamment : l'alignement et l'exploitation de la terminologie (*cf. figure 2*).

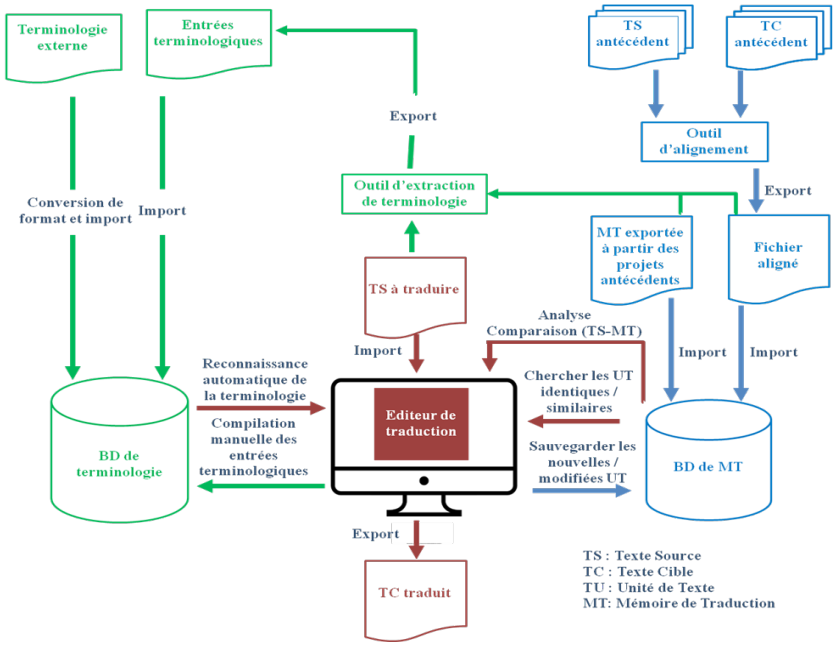

Figure 2 : Processus du système de la THAO

4.1. Alignement

L'alignement repose sur la comparaison des textes avec leur traduction, et à l'identification des correspondances d'un texte traduit avec sa version d'origine. Les algorithmes d'alignement se basent communément sur deux types de critères, formels ou linguistiques, pour morceler les textes en éléments, puis les apparier. D'autres indices permettent aussi de découper le texte en éléments, comme les signes de ponctuation. Le découpage peut être plus fin et reconnaître, grâce à des techniques d'analyse linguistiques, des éléments de plus petites longueurs (par exemple, des propositions ou des syntagmes nominaux).

4.2. Mémoire de traduction

Une mémoire de traduction est une base de données bilingue qui permet de stocker des segments du texte source avec leurs équivalents en langue cible. Elle est alimentée en permanence par le traducteur au fur et à mesure qu'il traduit. Ainsi, les éléments (segments de phrases, phrases) précédemment traduits seront proposés de façon automatique par le système lors de la traduction d'un nouveau texte contenant des passages similaires. Cependant, le traducteur peut les adopter, les réviser ou bien les écarter et traduire lui-même s'il n'est pas satisfait de la proposition.

4.3. Exploitation de la terminologie

Selon les systèmes de THAO, l'exploitation de la terminologie peut comprendre deux volets, à savoir la gestion et l'extraction terminologiques : La gestion terminologique consiste en l'élaboration d'une base terminologique. Elle offre un large spectre de fonctions, allant de la saisie et la personnalisation de fiches terminologiques jusqu'à la collaboration avec des experts du domaine. Cependant, l'extraction terminologique est conçue pour analyser des textes déjà traduits, d'un même domaine dans des langues différentes, et pour proposer des termes avec leur équivalent dans une autre langue.

5. Les outils de THAO présents sur le marché

Sur le marché, il existe plus de quatre-vingt outils de traduction assistée par ordinateur est d'aide à la traduction[3]. Devant une telle variété, il est difficile d'opter pour un outil ou un autre sans avoir recours à une liste de critères et de fonctionnalités à respecter.

5.1. Critères d'outils de THAO

Dans le souci de répondre aux besoins des traducteurs en fonction de leur habitude de travail, les outils de THAO sont devenus de plus en plus nombreux et variés dans leurs fonctionnalités et caractéristiques.

Ergonomie :

Afin de bénéficier de la cohérence des traductions et la diminution du temps en traduction, l'outil de THAO doit être configurable, pour permettre d'éliminer les paires de segments indésirables dans la mémoire de traduction et de ne garder que les segments pertinents. Cette configuration peut se faire à différents niveaux : supprimer ou modifier des segments, ou par le biais de la terminologie. Néanmoins, afin de réussir cette tâche, une bonne ergonomie est essentielle, assurant à l'outil une meilleure portabilité, la saisie automatique et la propagation des corrections.

3 https://www.atenao.com/blog/outils-tao/liste-des-logiciels-de-traduction-assistee-par-ordinateur/

Manipulation de mémoires de traduction :

Les mémoires de traduction doivent, généralement, enregistrer les travaux des traducteurs au fur et à mesure pour permettre les réutilisations ultérieures. Mais selon la nature de chaque travail, que les exigences de manipulation différent. Dans le cas d'un travail indépendant, les traducteurs souhaitent avoir une propagation automatique des modifications sur les textes à traduire. Cependant, dans le cas du travail collaboratif, où plusieurs traducteurs travaillent parallèlement sur le même projet, le partage de la mémoire et sa manipulation simultanée sont, en outre, privilégiés.

Recherche :

Tous les outils proposent des fonctions de recherche. Elles sont très pratiques pour trouver des phrases, des termes ou même des mots traduits auparavant. La recherche peut s'effectuer par simples chaînes de caractères ou par des expressions régulières. Les résultats de la recherche peuvent donner accès à la phrase contenant le mot recherché, au contexte du segment retrouvé ou au document tout entier. Plus ces options sont offertes plus le travail du traducteur est simplifié.

Compatibilité :

La plupart des outils de THAO exploitent et génèrent des ressources, telles que les données terminologiques, les mémoires de traduction et les règles de segmentation, en des formats bien déterminés. Sans aucune compatibilité entre ces outils ou les outils de gestion de contenu, l'exploitation de ces ressources reste limitée, voire impossible. Le recours à des convertisseurs pour pouvoir échanger les formats, exploiter ou modifier les ressources est nécessaire ; cependant, ces convertisseurs ne sont pas toujours disponibles. Il est donc essentiel que les outils puissent exporter et importer vers et à partir de formats standards.

Services associés :

En outre de leur fonction principale, les outils de THAO offrent des services supplémentaires, tels que l'alignement automatique et la gestion de la terminologie. Sans doute, ces services assurent aux traducteurs plus de flexibilité pour l'enrichissement des mémoires de traduction et les acquittent d'un fastidieux travail.

Etude comparative entre outils du THAO :

Afin d'appréhender certains outils à travers leurs différentes fonctionnalités et de cerner les pratiques et habitudes de la communauté des traducteurs à l'échelle internationale, nous nous sommes référencées à l'étude de la délégation générale à la langue française et aux langues de France (DGLFLF, 2015).

Cette étude a visé à mesurer l'adéquation entre les outils d'aide à la traduction et les besoins quotidiens des traducteurs. Elle se base sur une consultation publique menée auprès de 172 traducteurs en France au cours du mois de septembre et octobre 2014. Elle a concerné neuf outils, à savoir : Across[4], Business Translator[5], Enterprise server[6], MemoQ[7], MultiTrans[8], Similis[9], Trados[10], wordfast[11], et Déjavu[12].

Certes, l'échantillon exploité dans cette étude est limité à plusieurs niveaux, notamment le nombre des traducteurs qui y ont participé et la maîtrise des outils cités y compris leurs différentes versions, ainsi que la variation de langues sources et cibles. Mais, dans l'absence d'études similaires, cette dernière reste d'une grande importance pour cerner les outils les plus adéquats.

Outil	Utilisation	Ergonomie	Manipulation de MT	Recherche	Compatibilité	Classement général
Across	5	1	2	4	6	5
Business Translator	8	5	1	-	8	7
Enterprise server	8	5	1	-	8	7
MemoQ	2	1	3	2	2	1
Multi-Trans	6	4	6	5	7	8
Similis	7	3	1	1	5	4
Trados	1	2	3	6	1	2
Wordfast	3	1	5	3	3	3
Déjavu	4	2	4	6	4	6

Tableau 1 : Les résultats du classement des outils de la THAO

Les résultats obtenus dans cette étude sont repris dans le tableau ci-dessous. Les outils sont classés par ordre décroissant, où la valeur 1 indique le meilleur classement. Ce dernier est déduit à partir de la moyenne des notes attribuées par les différents traducteurs qui ont contribué à l'étude.

4 www.across.net
5 www.systransoft.com
6 www.systransoft.com
7 www.memoq.com
8 www.multicorpora.com
9 www.similis.org
10 www.translationzone.com
11 www.wordfast.net
12 www.atril.com

D'après ce tableau, il apparait que la solution Trados est la plus utilisée et que MemoQ vient en second pour le nombre de solutions installées. Par ailleurs, les solutions Across, MemoQ et w ordfast offrent une bonne ergonomie et permettent une bonne intégration dans l'espace de travail du traducteur. Par contre, pour la fonctionnalité de recherche et manipulation de la mémoire de traduction, ces solutions perdent leurs classements en tête de liste en faveur de la solution Similis. Cette dernière favorise, en outre, la recherche par expression régulière ; et privilégie plus les fonctions collaboratives tel que le travail simultané sur un projet. Concernant la compatibilité, la solution Trados gère au mieux les fonctions d'export et d'import des ressources. Par ailleurs, la solution MemoQ est appréciée pour sa gestion terminologique ainsi que sa possibilité de formatage similairement aux logiciels de traitement de texte habituels, qui facilite l'utilisation des fichiers tabulés ou des présentations.

En guise de conclusion, le tableau 1 confirme que l'efficience des solutions d'aide à la traduction varie d'une fonctionnalité à une autre. Néanmoins, la solution MemoQ vient en tête de liste dans le classement général par rapport aux outils objet de l'étude.

6. Conclusion

Cette étude a permis de faire émerger un certain nombre d'informations sur la qualité perçue par les traducteurs de certains outils ainsi que sur les usages de ces derniers par les traducteurs. Par ailleurs, nous pouvons déduire que plusieurs facteurs influencent la pertinence des outils de THAO notamment, le type de texte traité, la qualité des mémoires de traduction et surtout le temps.

Références

Pym A., Perekrestenko A., Starink B. (2006). « Translation Technology and its Teaching », Servei de Publicacions. Carrer de l'Escorxador s/n. 43003 Tarragona, Spain.

Quah C.K. (2006). Translation and technology. Basingstoke: Palgrave Macmillan.

Somers w .J., Somers H.L. (1992). An introduction to machine translation. London : Academic Press. Consultable en ligne : www.hutchinsweb.me.uk/IntroMT- TOC.html

Délégation générale à la langue française et aux langues de France-DGLFLF (2015). Mieux comprendre les outils d'aide à la traduction. Langues & Recherche.

Miftah N., Ataa Allah F., Taghbalout I. (2017). Sentence-aligned parallel corpus Amazigh-English, Actes de l'International Conference on Information and Communication Systems, 4-6 avril 2017, Irbid, Jordanie.

Taghbalout I., Ataa Allah F., El Marraki M. (2018). A Hybrid Approach for Amazigh-English Machine Translation, The 7th International Conference on Software Engineering and New Technologies "ICSENT'2018", 26-28 décembre 2018, Hammamet, Tunisie

Brace, C. (1994). Bonjour, Eurolang Optimizer, Language Industry Monitor, Issue Mar-Apr, http://www.lim.nl/monitor/optimizer.html

Annexe

Questionnaire

1. Quel est le nombre de traducteurs dans votre agence?
2. Quelles sont les langues que vous traduisez?
3. Quelles sont les langues les plus demandées ?
4. Est-ce que vous êtes une agence généraliste ou spécialisée ?

    a- Dans quel(s) domaine(s) votre agence de traduction est-elle spécialisée ? (traductions techniques, juridiques, médicales, financières.)

5. Est-ce que vous utilisez des outils de traduction assistée par ordinateur ?

    a- lesquels ?

    b- Suivez-vous des formations de l'exploitations d'outils de TAO ?

6. Quels sont les secteurs avec qui vous travaillez ?
7. Expérience professionnelle: depuis combien d'années exercez-vous ?

    a- 0-5, 5-10, 10-15, 15-20, 20-30, +30

8. Vos collaborateurs ont-ils un diplôme attestant de leur cursus académique de traducteur ?

    a- Etudes supérieures, études supérieures en traduction, autre

    b- Ont-ils suivi des formations en TAO au cours de leurs cursus.

TICAM'18, p. 75 - 83

La réalisation d'une base de données pour la reconnaissance audiovisuelle des chiffres amazighes

Ilham ADDARRAZI[1], Hassan SATORI[2], Khalid SATORI[1]
[1] Laboratoire Informatique, Imagerie & Analyse Numérique (LIIAN)
Faculté des Sciences Dhar El Mahraz - Université Sidi Mohamed Ben Abdellah - Fès, Maroc
{ilham.adrz, khalidsatori}@gmail.com
[2] Intelligence Artificielle Systèmes Complexes et Modélisation (IASCM)
Département de mathématiques et d'informatique - Faculté Pluridisciplinaire Nador
Université Mohammed Premier
hsatori@yahoo.com

Résumé

La réalisation d'un système de reconnaissance audio-visuelle permet d'augmenter les performances de système de reconnaissance de parole surtout dans des milieux bruitées. D'ailleurs, l'ajout de l'information visuelle à travers d'image vidéo du locuteur nécessite l'existence d'un corpus de données audiovisuelle. L'objectif de ce travail consiste à réaliser une première base de données audiovisuelle pour la langue amazighe nommée AmDigit_AVSR. Ce corpus contient des enregistrements vidéo de 30 locuteurs (15 femmes, 15 hommes) qui ont prononcé les dix chiffres amazighes.

Mots-clés : Reconnaissance audio-visuelle, RAP, Amazighe.

Introduction

La Reconnaissance Automatique de la Parole (RAP) est un système permettant d'interpréter une langue naturelle humaine sur une machine. Sa structure est constituée de l'ensemble des ressources pour transformer le signal acoustique en séquences des mots correspondantes à celles prononcées par un locuteur. Cependant, l'utilisation de ces systèmes dans les conditions d'enregistrement bruitées entraine un abaissement des performances du système RAP (Le Prell *et al.*, 2017). L'utilisation d'information conjointement au signal de parole est une voie classique pour améliorer les performances et la robustesse des systèmes de reconnaissance automatique de la parole. De nombreux travaux sur la perception de la parole ayant montré l'importance des informations visuelles dans le processus de reconnaissance chez l'homme (McGurk et MacDonald, 1976)

En particulier, l'exploitation des informations visuelles tel le mouvement des lèvres du locuteur, qui est une méthode encourageante pour la reconnaissance de la parole en milieux bruités. Ces informations sont intégrées aux informations acoustiques pour donner un système capable de reconnaissance audiovisuelle de la parole (RAVP). Partant, le RAVP est utilisé pour améliorer le taux de reconnaissance de la parole surtout dans un milieu bruité (Biswas *et al.*, 2015 ; Makhlouf *et al.*, 2016).

Pour tester la performance du système RAVP, la création d'une base de données audiovisuelle est nécessaire. Toutefois, la réalisation de ce corpus nécessite des données audio et vidéo qui ont été enregistrées simultanément.

Au cours des dernières années, un effort a été fait pour créer des bases de données pour la communauté de la recherche audiovisuelle. Le premier corpus a été proposé par Petajan *et al.* (1984) pour réaliser un système de lecture labial capable de reconnaitre les dix digits et les alphabets. Pigeon *et al.* (1997) ont présenté (M2VTS). Il s'agit d'enregistrements audio et de séquences vidéo de 37 sujets prononçant des chiffres de 0 à 9 en cinq séances. Messer *et al.* (1999) ont étendu M2VTS et l'ont désigné par (XM2VTS). XM2VTS est créé pour étendre la taille de l'échantillon de M2VTS à 295 locuteurs. Les autres bases de données sont proposées dans la littérature comme Tulips1, CUAVE (Patterson *et al.*, 2002), BANCA (Bailliere *et al.*, 2003), etc.

La plupart des bases de données RAVP précédentes sont en anglais ou dans une autre langue (Richie *et al.*, 2009 ; Benezeth *et al.*, 2011). Cependant, le travail proposé dans cet article concerne la réalisation d'une base de données audiovisuelle pour la langue amazighe. À notre connaissance, ce corpus est la première base de données utilisant cette langue marocaine.

Notre article est organisé comme suit. Nous présentons la langue utilisée pour notre base de données audiovisuelle dans la section 2. Dans la section 3, nous décrivons les détails de notre corpus. La section 4 illustre les systèmes qui utilise cette base de données. La section 5 conclut cet article et présente les futures de ce travail.

1. La langue de la base de données RAVP

La langue amazighe connue sous le nom de berbère ou tamazight est l'une des plus anciennes langues. Elle est parlée dans l'Afrique du Nord et Sahara-Sahel.

Au Maroc, 28 % de populations utilisent l'amazighe. Ils sont regroupés en trois principales variétés régionales : Tarifit parlé au nord, Tamazight au Maroc central et au Sud-est, et Tachelhit parlé au sud.

Vu l'importance de la langue berbère, l'Institut Royal de la Culture Amazighe (IRCAM) a proposé en 2003 un système graphique de la langue berbère. Ce système s'appelle Tifinaghe-IRCAM. Il est considéré comme un système officiel au Maroc pour écrire l'amazighe.

Tifinaghe-IRCAM contient (Ataa Allah et Boulaknadel, 2010) :

- 27 consonnes : les labiales (ⵃ, ⵀ, ⴼ), les dentales (ⵜ, ⵟ, ⴷ, ⴹ, ⵍ, ⵏ, ⵇ, ⵅ), les alvéolaires (ⵣ, ⵙ, ⵚ, ⵛ), les palatales (ⵊ, ⵉ), les vélaires (ⴽ, ⵅ), les labiovélaires (ⴽⵡ, ⵅⵡ), les uvulaires (ⵖ, ⵅ, ⵄ), les pharyngales (ⵄ, ⵃ) et la laryngale (ⵀ);
- 2 semi-consonnes : ⵢ et ⵡ ;
- 4 voyelles : trois voyelles pleines ⴰ, ⵔ, ⵓ et la voyelle neutre (ou schwa) ⴻ qui a un statut assez particulier en phonologie amazighe.

Les syllabes confirmées en langue amazighe sont : V, CV, VC, CVC, C, CC et CCC où V indique une voyelle tandis que C indique une consonne (Ridouane, 2003).

Le tableau 1 montre les dix chiffres amazighes utilisés dans la base de données proposée :

Digits	Transcription en anglais	Transcription en arabe	Transcription en Tifinaghe	Syllables
0	AMYA	اميا	ⵄⵎⵉⴰ	VC-CV
1	YEN	يان	ⵉⴰⵏ	CVC
2	SIN	سين	ⵙⵉⵏ	CVC
3	KRAD	كراض	ⴽⵔⴰⴹ	VC-CVC
4	KOZ	كوز	ⴽⵓⵣ	CVC
5	SMMUS	سموس	ⵙⵎⵎⵓⵙ	CCV-VC
6	SDES	سضيس	ⵙⴹⵉⵙ	CCVC
7	SA	سا	ⵙⴰ	CV
8	TAM	تام	ⵜⴰⵎ	CVC
9	TZA	تزا	ⵜⵣⴰ	CC-CV

Tableau 1 : Présentation des chiffres utilisés avec une transcription en anglais, en arabe et en Tifinaghe

2. Le corpus AM_AVSRDigit

La base de données AmDigit_AVSR (Amazighe Digit Audio-visual speech recognition) est un corpus de langue amazighe. Elle est développée au laboratoire IASCM à la faculté pluridisciplinaire Nador (FPN) dans le contexte de préparer et de tester un système de reconnaissance audiovisuelle de parole pour les chiffres amazighes (Addarrazi *et al.*, 2017). D'ailleurs, ce corpus contient des enregistrements vidéo de 30 locuteurs en prononçant les dix chiffres (voir table 1) dans un milieu non bruité.

2.1. Les locuteurs

AmDigit_AVSR corpus contient 30 locuteurs (15 femmes et 15 hommes) qui sont âgés entre 18 et 60 ans. Chaque personne énonce dix fois une suite de dix chiffres amazighes dans le but d'avoir une bonne connaissance de chiffre dans les différents états quelques soit le locuteur et son style de prononciation.

Cette base de données comprend les interlocuteurs de différents genres (sans ou avec barbe, sans ou avec foulard).

Locuteur	Femme	Homme
Nombre de locuteurs	15	15
Nombre de répétions	10	10
Nombre de données	1500	1500

Tableau 2 : Le contenu de l'AmDigit_AVSR

2.2. La configuration du corpus

Les videos dans AmDigit_AVSR sont enregistrés à 25 images/s en utilisant une camera d'une résolution de 1280 * 720 située à 70 cm de locuteur (voir figure 2). L'image capturée par les videos contient le visage frontal du locuteur et sa partie supérieure du corps avec un arrière-plan simple. Le tableau 2 montre plus de détails sur le corpus :

Paramètre	Valeur
Echantillonnage	16 KHz
Résolution	720*1280
Format de fichier audio	.wav
Image par seconde	25
Corpus	10 chiffres amazighes
Nombre de répétitions par chiffre	10
L'environnement	Sans bruit

Tableau 3 : Paramètres d'enregistrement utilisés pour la réalisation du corpus amazighe digits

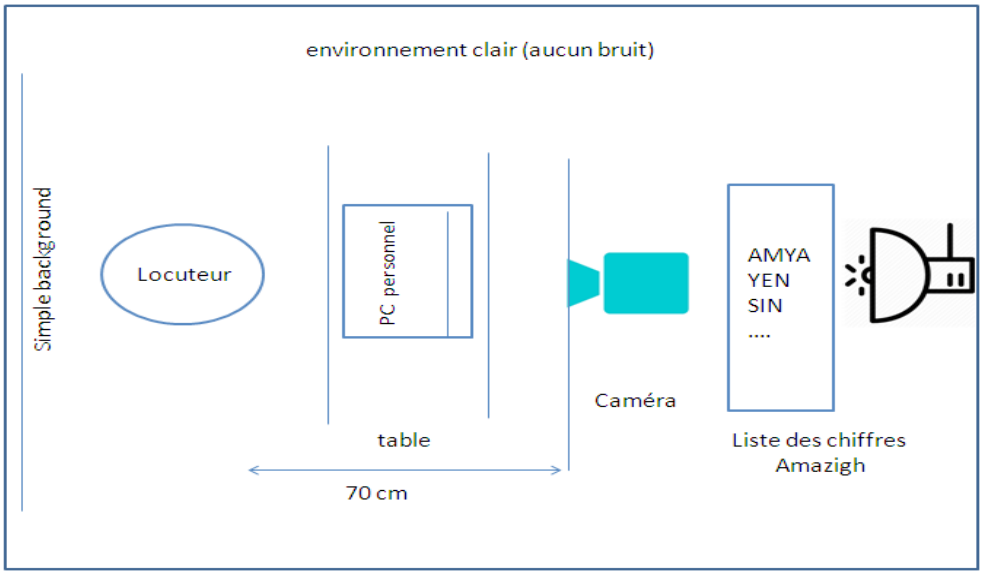

Figure 1 : Environnement d'enregistrement

3. L'étiquetage et préparation des vidéos

3.1. Séparation des données du corpus

Les sessions d'enregistrement commencent en demandant aux locuteurs de répéter dix fois chaque chiffre amazighe devant une caméra. En cas d'erreur, l'intervenant peut répéter le chiffre.

La première opération consiste à convertir les séquences vidéo enregistrées de l'extension « mp4 » vers l'extension « .avi ». À partir du fichier video, un script automatique est utilisé pour extraire un flux audio sous forme un signal de l'extension « .wav ».

Sous audacity, les séquences vidéo sont analysées pour les segmenter manuellement en se basant sur la durée de prononciation de chaque chiffre.

Figure 2 : Les dix répétitions d'AMYA

3.2. Organisation des données audiovisuelles

Les enregistrements sont sauvegardés sous une arborescence bien spécifique comme indiqué dans la figure 3. Les fichiers audio extraits sont placés dans un répertoire appelé WAV et les vidéo dans AVI. Nous avons créé pour chaque locuteur un dossier nommé par les deux premiers alphabets de son nom et de son prénom. Pour L100.avi, L1 indique le nom de l'intervenant, 0 signifie le chiffre AMYA et 0 signifier le numéro de répétition.

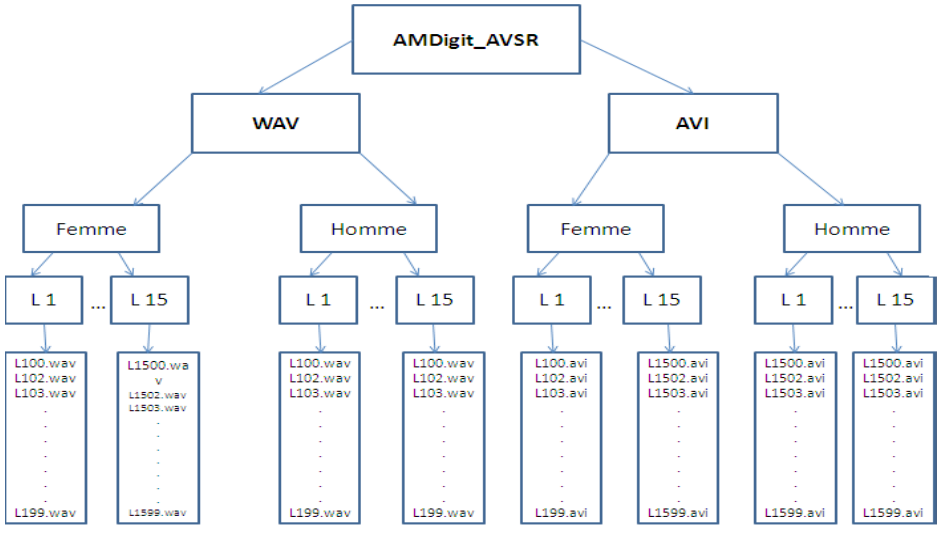

Figure 3 : L'arborocence d'AmDigit_AVSR

4. Evaluation

La structure du système de reconnaissance audiovisuelle pour les chiffres amazighes (Addarrazi *et al.*, 2017) est présentée dans la figure ci-dessous. Ce système possède trois modules qui sont : le module de reconnaissance visuelle, le module de reconnaissance acoustique et le module d'intégration. La performance de chaque modèle nécessite une préparation d'AMDigit_AVSR.

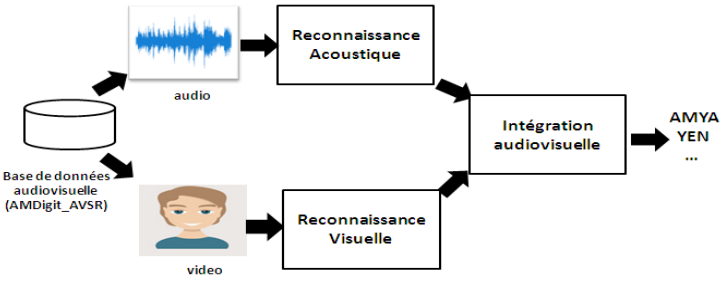

Figure 4 : Les modules de RAVP

4.1. La détection visuelle

Dans le but d'évaluer les performances de notre système de reconnaissance audiovisuelle, la dernière étape de préparation du corpus consiste d'extraire les frames à partir de chaque vidéo segmentée. Ces images sont utiles pour faire la détection des lèvres de chaque locuteur.

Extraire les caractéristiques du visage est une étape nécessaire dans les systèmes RAVP. La détection de visages est le fait de trouver l'emplacement de visage dans une image ou une vidéo. Pour ce faire, l'algorithme Viola-Jones est utilisé. L'application de cette approche sur les images de notre corpus donne un résultat satisfait (99%).

Une fois le visage est isolé, il est recommandé de localiser la région d'intérêt (la bouche). Le mouvement de lèvres d'un locuteur occupe une place très privilégiée dans notre système. Il offre une source visuelle d'information pour la reconnaissance de la parole. Cette zone d'intérêt est localisée en utilisant toujours l'algorithme Viola-Jones. Le résultat de la méthode est un rectangle autour de la bouche détectée (96.6%).

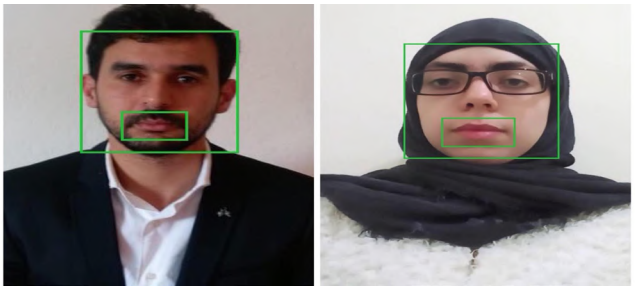

Figure 5 : Les résultats des détections de visage et la bouche par Viola-Jones

4.2. La reconnaissance acoustique

Le développement de notre RAP (Satori et ElHaoussi, 2014) dépend des audio extraits du corpus AMDigit_AVSR. C' est un système de reconnaissance automatique de la parole pour les chiffres amazighes. Ce système, comme montré dans la figure 6 comporte typiquement 5 modules :

- l'extraction des caractères acoustique, qui va extraire des séquences de paramètres acoustiques ;
- le modèle de prononciation, qui associe les mots à leur description phonétique ;
- le modèle acoustique, qui va décrire la distribution des phonèmes ;
- le modèle linguistique, qui va présenter pour un mot donnée la séquence de mots la plus probable ;
- le décodeur, qui va traiter les information reçues depuis le modèle acoustique, il les combinent avec le modèle linguistique pour donner un résultat.

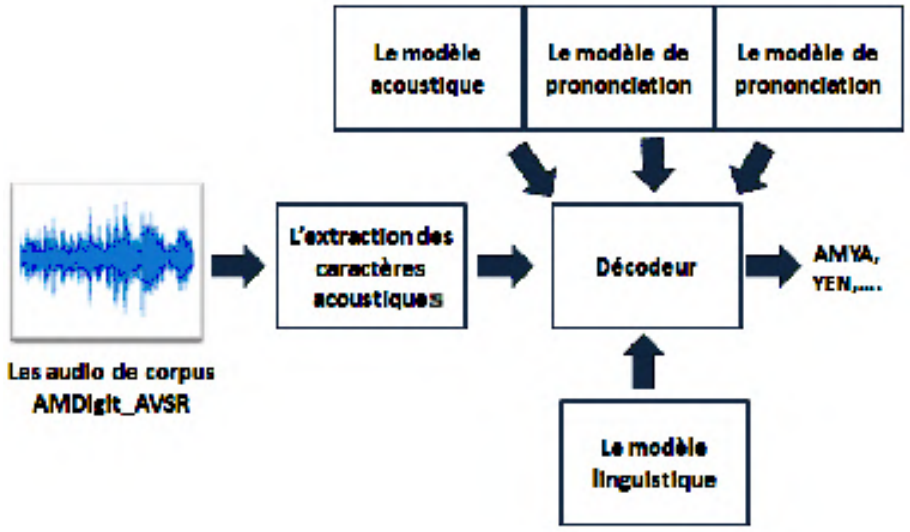

Figure 6 : L'architecture du système de reconnaissance de la parole

La performance de notre système de reconnaissance est fortement liée aux données de corpus proposé. Il a obtenu la meilleure performance de 92,89%.

5. Conclusion

L'objectif principal de ce travail est la mise en œuvre d'un corpus des données audiovisuelles pour élaborer et tester un système de reconnaissance audiovisuelle des chiffres amazighes. Cette base de données est le premier corpus utilisant la langue amazighe. Elle contient 1500 fichiers vidéo et audio de 30 locuteurs qui ont énoncé les 10 premiers chiffres amazighes dans un milieu non bruité.

Nos efforts restent concentrés sur l'augmentation de la taille de notre base de données (les autres chiffres, les alphabets, les phrases, ...) prononcées par plusieurs types de locuteurs dans des milieux différents afin de tester la fiabilité réelle de notre système de reconnaissance.

Références

Addarrazi, I., Satori, H., Satori, K. (2017). Amazigh audiovisual speech recognition system design. Intelligent Systems and Computer Vision (ISCV). pp. 1-5.

Ataa Allah, F., Boulaknadel, S. (2010). Amazigh Search Engine: Tifinaghe Character Based Approach. In Proceeding of the International Conference on Information and Knowledge Engineering. pp. 255-259.

Bailly-Bailliére, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Ruiz, B. (2003). The BANCA database and evaluation protocol. In International conference on Audio-and video-based biometric person authentication. pp. 625-638. Springer, Berlin, Heidelberg.

Benezeth, Y., Bachman, G., Le-Jan, G., Souviraà-Labastie, N., Bimbot, F. (2011). BL-Database: A French audiovisual database for speech driven lip animation systems (Doctoral dissertation, INRIA).

Biswas, A., Sahu, P. K., Chandra, M. (2015). Multiple camera in car audio–visual speech recognition using phonetic and visemic information. Computers & Electrical Engineering, Vol. 47, pp.35-50.

Le Prell, C. G., Clavier, O. H. (2017). Effects of noise on speech recognition: challenges for communication by service members. Hearing research. Vol. 349, pp. 76-89.

Makhlouf, A., Lazli, L., Bensaker, B. (2016). Evolutionary structure of hidden Markov models for audio-visual Arabic speech recognition. International Journal of Signal and Imaging Systems Engineering, 9(1):55-66.

McGurk, H., MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264(5588), 746.

Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G. (1999). XM2VTSDB: The extended M2VTS database. In Second international conference on audio and video-based biometric person authentication. Vol. 964, pp. 965-966.

Patterson, E. K., Gurbuz, S., Tufekci, Z., Gowdy, J. N. (2002). CUAVE: A new audio-visual database for multimodal human-computer interface research. In Acoustics Speech and Signal Processing (ICASSP). Vol. 2.

Petajan, E. D. (1984). Automatic lipreading to enhance speech recognition (speech reading).

Pigeon, S., Vandendorpe, L. (1997). The M2VTS multimodal face database (release 1.00). In International Conference on Audio-and Video-Based Biometric Person Authentication. pp. 403-409. Springer, Berlin, Heidelberg.

Richie, S., w arburton, C., Carter, M. (2009). Audiovisual database of spoken American English. Linguistic Data Consortium.

Ridouane, R. (2003). Suites de consonnes en berbère: phonétique et phonologie. Doctoral dissertation, Université de la Sorbonne nouvelle-Paris III.

Satori, H., ElHaoussi, F. (2014). Investigation Amazigh speech recognition using CMU tools. International Journal of Speech Technology. 17(3): 235-243.

TICAM'18, p. 85 - 93

Improving English-Arabic statistical machine translation using the linguistic knowledge on data

Safae BERRICHI, Azzeddine MAZROUI
Department of computer Science
Faculty of Sciences, University Mohamed First - Oujda, Morocco
{berrichi.safae,azze.mazroui}@gmail.com

Abstract

Machine translation between two languages based on statistical methods has made significant progress during these recent years. However, despite the good performances achieved by these methods where the source and target languages are morphologically close. These performances remain below expectations for two languages with very different morphological structures, as in the case of Arabic and English. To overcome the challenges facing statistical machine translation (SMT) between Arabic and English, we introduce in this paper a new approach that exploits morphosyntactic information in the alignment phase between these two languages. This has reduced the error rate of word alignment, and consequently improved the machine translation quality.

Keywords : Statistical Machine Translation, Phrase-Based translation, Parallel corpus, Alignment corpus, Morphological Analysis.

1. Introduction

Machine translation occupies a preponderant position in the field of natural language processing (NLP). The approaches developed in the machine translation field can be classified into three families (Okpor, 2014; Kituku, Muchemi and Nganga, 2016). The first family takes in rule-based approaches that use bilingual dictionaries and a set of linguistic rules (morphological, syntactic, and semantic). The second family that is prevalent in the state of art concerns statistical approaches that are based on probabilistic representations. This allows to model both the association of word sequences between source and target languages, and the generation of a sentence grammatically correct in target language. These approaches use parallel corpora for each language pair. Finally, the third family consists of hybrid approaches that combine linguistic rules with statistical approaches. These methods have attracted increasing interest in recent years.

The automatic translation between Arabic and English is a difficult problem and poses several challenges. This is due to the nature of the Arabic language morphology, which is relatively complex and different from that of the English language. Indeed, the phenomenon of agglutination, which is very prevalent in Arabic, makes that an Arabic word can correspond to a whole sentence in English. Thus, the word "وسيعطيها" corresponds to five words in

English “and he will give her”. This complexity has a negative impact on the alignment phase quality. This leads to a decrease in the performance of the SMT systems, which all use an alignment phase.

In this work, we sought to improve the English-Arabic translation through the incorporation of morphosyntactic information (stem, lemma) into the learning data. we evaluated the contributions of these information in the alignment phase and in the translation system.

In the next section, we will present a description of the SMT process based on phrases (Koehn *et al.*, 2003). we then review, in section 3, some previous work. In section 4, we describe our approach, and we present the used tools and resources. Section 5 will be devoted to the evaluation phase. At the end, the paper will be concluded with some perspectives.

2. Phrase-based Statistical Machine Translation

Phrase-based statistical machine translation is carried out in two basic steps: alignment and translation.

In the alignment phase, a parallel corpus is used to deduce the word alignments of the corpus sentences. These alignments are then used to extract phrases estimated by a translation probability, which allows to generate the translation model. Finally, the language model is developed using a monolingual corpus of the target language.

During the translation process, the input source sentence is first segmented into phrases, and then translated into several hypotheses learned in the training pair sentences. The selected translation will be the one that maximizes by the decoding mechanism the probabilities of the translation model and that of the language.

3. State Of the Art

Several researchers have focused on the challenges in machine translation related to the Arabic language. Thus, (Soudi and Farghaly, 2012) presented a comprehensive study of the challenges encountered in this area. Indeed, they pointed out that the lack of diacritic marks in the Arabic text, the ambiguity in the analysis of missing information, the existence of synonyms for Arabic nouns or expressions that have no equivalent in English are part of the morphological generalization challenges. They then proposed solutions to these challenges. Similarly, (Alkhatib and Shaalan, 2018) have recalled some challenges encountered in different variants of Arabic: classical Arabic, standard Arabic and dialectal Arabic. They thus raised the problems related to the conceptualization of word sense disambiguation, to the metaphor and the Named Entity Recognition that generate several linguistic problems in machine translation.

Most of researches in machine translation related to the Arabic language have used the hybrid approach that combines statistical methods and linguistic rules. The majority of them are interested in translating between Arabic and English. In order to improve the quality of translation, these researches have made changes in the pretreatment and post-processing phases for both directions of parallel corpora. Thus, (Mallek *et al.*, 2018) were interested in the problems encountered in the Arabic-English translation of the tweets. They noted that the

normalization and segmentation of words have improved the blue score by 4 points. Similarly, (Khemakhem *et al.*, 2015) and (Ghaffar and Fakhr, 2011) carried out a morphological analysis that measures the impact on the blue score of the Arabic word segmentation and the preprocessing of numbers, dates and the grouping of the proper names. Finally, (El Kholy and Habash, 2012) have compared several tokenization schemes and concluded that the best performer is Arabic Treebank (ATB). The system was tested on English-Arabic parallel corpus gathered from the Linguistic Data Consortium[1] (LDC). All these works used the MADA analyzer in the morphological analysis phase (Habash and Rambow, 2005).

4. Description of the approach

w e get started this section by presenting the resources and tools used in the developing of the machine translation system. Then, we give a description of our method.

4.1. Data & tools

Improving the statistical machine translation quality based on phrases depends on the availability of a large linguistic dataset. w e used in this work data that we collected from the English-Arabic United Nation parallel corpus (Ziemski *et al.*, 2016). w e have subdivided these data into three sets: the training set used to build the statistical models (the translation model and the language model), the development set that maximizes the linear combination of models and weights for a better translation, and the test set for the evaluation phase of the SMT developed system. Table 1 illustrates the statistics about these three sets.

	Number of sentences	***Number of words***		***Vocabulary***	
		English	***Arabic***	***English***	***Arabic***
Training set	5597	165687	148035	21184	10552
Development set	1000	30490	27355	7716	4533
Test set	1400	42050	37748	9378	5406

Table 1: Detailed statistics on the corpus

w e used these sets in two experiments in the popular Moses-based translation system[2] (Koehn *et al.*, 2007). The first experiment, noted Baseline, does not apply morphological preprocessing, while the second enriches all the data by morphosyntactic information. For the Arabic language, these characteristics are obtained by using the morphological disambiguation system of the Arabic language ALKHALIL (Boudchiche and Mazroui, 2016); and for the English language, we used the Stanford analyzer (Manning *et al.*, 2014). These two systems provide for each analyzed word its morphological tags such as the lemma and the stem.

1 https://www.ldc.upenn.edu/
2 http://www.statmt.org/moses/?n=Moses.Baseline

In the word alignment phase, the GIZA++[3] tool (Och and Ney, 2003), which implements IBM1-5 models and HMM, was used to obtain word alignment from a parallel corpus. we were interested in this work by the models of IBM1-2 and HMM.

4.2. Description of the method

The different steps of the word alignment and translation process are outlined in Figure 1.

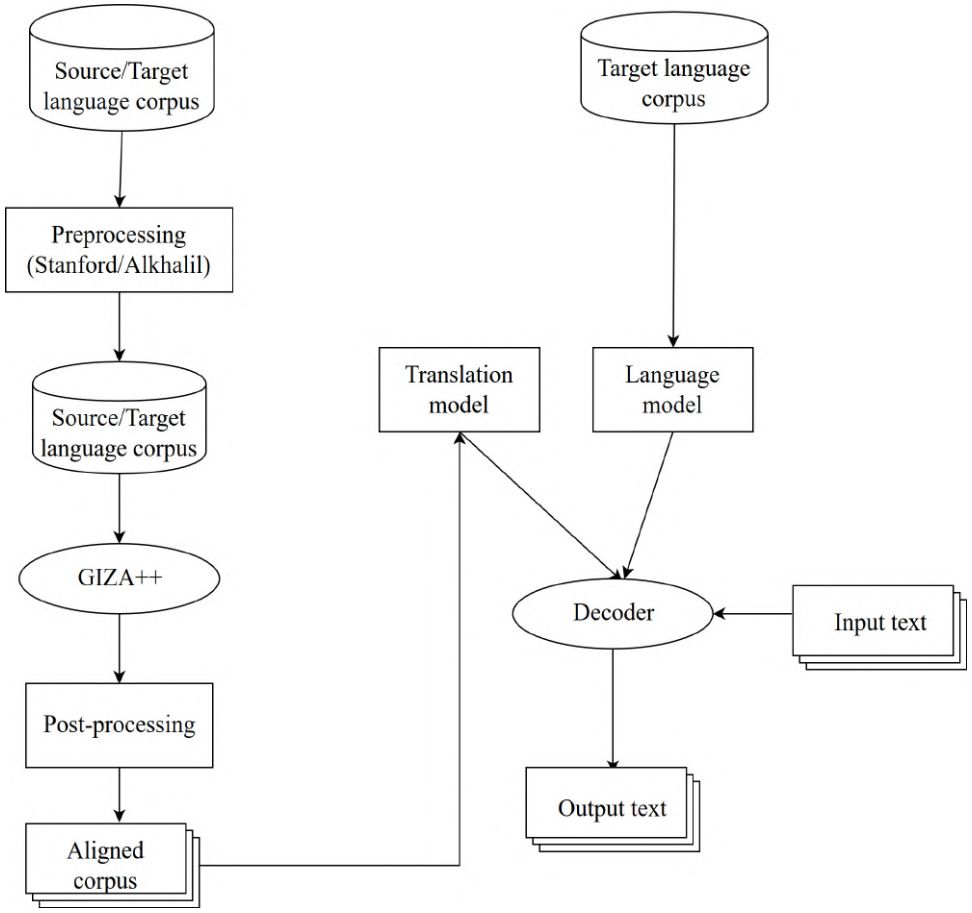

Figure 1: Word alignment and machine translation process

3 http://www.fjoch.com/giza-training-of-statistical-translation-models.html

w e will give bellow a description of each step.

1. Preprocessing: this step consists in enriching the words of the parallel corpus by morphosyntactic tags using Alkhalil for Arabic words and Stanford for English words. The first preprocessing is to segment the words into proclitic+stem+enclitic. In the second preprocessing, we extract the word lemmas to obtain the proclitic+lemma+enclitic segmentation. w e, thus, obtain two representations of the corpus, one based on the stems and the other on the lemmas.
2. Alignment: we proceed in this step to the words alignment of one of the parallel corpus representations by the GIZA++ tool. After aligning words in the representation of the corpus based on lemmas, we use a post-processing that replaces the lemmas by their corresponding stems. The alignment is achieved in both Arabic-English and English-Arabic directions.
3. Translation: the translation step takes as input the parallel corpus and the output of the alignment phase in both directions to build the probabilistic models (translation model and language model). In order to produce the best translation of the input text, the system thereafter applies the decoding phase on these models to calculate the likelihood of the translation hypotheses in the target language.

5. Word Alignment results

In order to measure the impact of morphosyntactic preprocessing on alignment quality, we performed three types alignment in both directions English-to-Arabic and Arabic-to-English:

- Baseline-Alignment is a standard system that proceeds directly to the text alignment at the word level without any preprocessing.
- Stem-Alignment performs at the beginning a preprocessing on the data to segment the word into proclitic, stem and enclitic. Then, the system seeks to align the word segments instead of aligning the words.
- Lemma-Alignment is similar to the previous one but replacing the stem with the lemma. After alignment, the lemmas are replaced by their corresponding stems.

The training data of the alignment phase are similar to those of the translation phase, and they consist of 5579 pairs of sentences (see Table 1). Since a large aligned Arabic-English corpus is not available, we invited two translation experts to manually align 100 randomly extracted sentences from the training corpus, for using them in the evaluation of the alignment phase.

For word alignment measurements, we used the precision, recall, F-measure, and alignment error rate (AER) described in (Och and Ney, 2003). These measures are defined by the following formulas:

$$precision = \frac{card\,(A_S \cap A_A)}{card\,(A_S)} \qquad F\text{-}measure = 2\,\frac{precision * recall}{precision + recall}$$

$$recall = \frac{card\,(A_S \cap A_A)}{card\,(A_A)} \qquad AER\,(S, P, A) = 1 - \frac{card\,(A_S \cap A_A)}{card\,(A_A) + card\,(A_S)}$$

where

A_S = Set of all alignments proposed by the system.

A_A = Set of all the alignments suggested by the annotators.

we, therefore, calculated these measurements on the test set consisting of 100 sentences aligned by the two experts for each of the three approaches used in the word alignment phase. The results obtained are presented in Table 2.

	Precision	**Recall**	**F-measure**	**AER**
Baseline-Alignment				
English-Arabic	52.81	64.71	58.16	70.91
Arabic-English	61.12	65.28	63.13	68.43
Stem-Alignment				
English-Arabic	59.49	73.38	65.71	67.14
Arabic-English	70.76	77.54	73.99	63.00
Lemma-Alignment				
English-Arabic	61.03	73.80	**66.81**	66.59
Arabic-English	72.17	78.13	**75.03**	62.48

Table 2 : Alignment phase evaluation

The results obtained show that the Baseline-Alignment, which does not perform any preprocessing on the data, is the least efficient. Indeed, its F-measure is lower by almost 10% compared to the other two alignments types. In addition, the use of lemmas instead of stems in word segmentation has achieved an improvement in the alignment quality. Finally, alignment in the Arabic-English direction is more efficient for the three alignment types than in the English-Arabic direction.

4.3. MT results

After evaluating the alignment quality of the three approaches used in the GIZA ++ system, we assessed the impact of these approaches on the Moses translation engine. Indeed, alignment models were used as the starting point for the phrase-based statistical machine translation.

Then, we calculated the measure Blue Score (Papineni *et al.*, 2002) to evaluate the translation quality. The evaluation was carried out on the test corpus containing 1400 sentences as well as on the development corpus, which contains 1000 sentences (See Table 1). we present, in Table 3, the results obtained for each alignment approach choice.

Blue Score	**dev**	**Test**
Baseline-Alignment		
English-Arabic	22.12	**21.29**
Stem-Alignment		
English-Arabic	35.32	**34.58**
Lemma-Alignment		
English-Arabic	35.69	**35.11**

Table 3: Translation phase evaluation for English-Arabic direction

It is clear from these results that words' segmentation in clitics plus stem or lemma makes it possible to considerably improve the translation quality. Indeed, the bleu score has increased from 21.29 for the approach without segmentation (Baseline-Alignment) to more than 34 for segmentation approaches. Similarly, the use of lemmas instead of stems in the segmentation has contributed to improving translation quality.

5. Conclusion and future works

In this paper, we have proposed two approaches that rely on linguistic enrichment in aligned sentence pairs. They consist in segmenting the words on their canonical forms (stem or lemma). The impact of these approaches has been evaluated initially in the alignment phase and later in the translation phase. The test results obtained showed the significant contribution of the word segmentation phase on the quality of alignment and translation quality. In addition, it is clear from the tests that the performances of the statistical machine translation system are dependent on the accuracy of the alignment phase. As future work, we plan to make the alignment system more efficient and to define the alternative translation models that incorporate morphological information useful in the translation phase.

References

Alkhatib, M., Shaalan, K. (2018). The Key Challenges for Arabic Machine Translation. In Shaalan, K., Hassanien, A. E., and Tolba, F. (eds) *Intelligent Natural Language Processing: Trends and Applications*. Cham: Springer International Publishing, pp. 139-156. doi: 10.1007/978-3-319-67056-0_8.

Boudchiche, M., Mazroui, A. (2016). Approche hybride pour le développement d'un lemmatiseur pour la langue arabe. In Proceedings of the 13th African Conference on Research in Computer Science and Applied Mathematics. CARI'16. Hammamet, Tunisia. October 10-14.

El Kholy, A., Habash, N. (2012). Orthographic and morphological processing for English–Arabic statistical machine translation. *Machine Translation*. 26(1–2):25-45.

Ghaffar, S. A., Fakhr, M. w. (2011). English to Arabic Statistical Machine Translation System Improvements using Preprocessing and Arabic Morphology Analysis. In Recent Researches in Mathematical Methods in Electrical Engineering and Computer Science. pp. 50-54.

Habash, N., Rambow, O. (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In Association for Computational Linguistics. pp. 573–580. doi: 10.3115/1219840.1219911.

Khemakhem, I. T., Jamoussi, S., Hamadou, A. B. (2015). Arabic-English Semantic w ord Class Alignment to Improve Statistical Machine Translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*. pp. 663-671.

Kituku, B., Muchemi, L., Nganga, w . (2016). A Review on Machine Translation Approaches. *Indonesian Journal of Electrical Engineering and Computer Science*, 1(1):182-190.

Koehn, P., Hoang, H., Birch, A., (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics. pp. 177-180.

Koehn, P., Och, F. J., Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, Association for Computational Linguistics, pp. 48-54.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., McClosky. D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. pp. 55–60.

Och, F. J., Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*. 29(1):19-51.

Okpor, M. D. (2014). Machine translation approaches: issues and challenges. *International Journal of Computer Science Issues (IJCSI).* 11(5):159.

Papineni, K., Roukos, S., w ard, T., Zhu, w . (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics. pp. 311–318.

Soudi, A., Farghaly, A. (2012). *Challenges for Arabic machine translation*. John Benjamins Publishing.

Ziemski, M., Junczys-Dowmunt, M., Pouliquen, B. (2016). The United Nations Parallel Corpus v1. 0. In Proceedings of LREC, *Portoroz, Slovenia*, May 23-28.

TICAM'18, p. 95 - 103

Arabic sentiment classification using POS Tagger and SVM

Ibtissam Touahri, Azzeddine Mazroui
Department of Computer Science Faculty of Sciences, University Mohamed First Oujda, Morocco
{ibtissamtouahri555, azze.mazroui}@gmail.com

Abstract

Sentiment Analysis (SA) helps to automatically extract emotions and feelings from comments available on social media. The two main families of approaches used for sentiment analysis are the supervised approaches that are based on annotated corpus and consist in generating a model, and the unsupervised ones that are based solely on a lexicon. we have developed in this article a method to analyze feelings for the Arabic language. In order to remedy the problem of vector size representing the data, which is proportional to the lexicon size, we used a POS Tagger to divide the lexicon according to the word tags. By partitioning the lexicon, we minimized the length of the vector representing the data, and as a result we reduced the execution time. we also improved the accuracy of the analysis by considering the word negation context.

Keywords: Arabic language, Sentiment Analysis, Supervised approach, POS Tagger.

1. Introduction

In order to determine customer preferences in marketing domain, the quality of service at a restaurant or even the prediction of the candidate who will be elected in elections, we can extract and analyze opinions from social media websites and micro blogs containing comments that carry information on one or more specific topics. In Arabic, a person can express an opinion using either modern standard Arabic (MSA) or dialectal Arabic (DA) or a combination of both. MSA is widely used in newspapers and magazines, while DA is dominant in social media. The MSA faces many challenges among them, the absence of diacritical marks that makes the word ambiguous because it can be interpreted in many ways. For instance, the word "مرض" can be read as "مُرْضٍ" (satisfying) that expresses a positive feeling or "مَرَضٌ" (sickness) that refers to a negative situation. DA also has its own complexities. It varies from one region to another, so a word can be expressed in many ways.

The first step in sentiment analysis is the detection of subjectivity that helps to determine whether a sentence is a fact or an opinion (Abdul-Mageed *et al.*, 2014). A fact is objective and can be proved, whereas an opinion is subjective and is based on feelings. The second step in sentiment analysis field is to determine the polarity of the subjective sentences. The simplest form of polarity is to consider it either positive, negative or neutral (ElSahar and El-Beltagy, 2015). It can be also fine-grained, in other words, it can be more precise about the level and considered as very positive, positive, neutral, negative or very negative .

The two main families of approaches used for sentiment classification are the supervised approaches that are based on a model built from an annotated corpus, and the unsupervised approaches that use only lexicons to predict feelings (Abdulla *et al.*, 2014). In most cases, the use of supervised methods with a domain-specific lexicon makes it possible to obtain better results but remains expensive in term of time because it requires the annotation of a sentiment corpus. In the family of supervised classifiers, several studies figure out that Support Vector Machine (SVM) surpasses Bayesian classifiers (NB) (Abuelenin *et al.*, 2018), (Abdulla *et al.*, 2013) and k-nearest neighbors (k-NN) (ElSahar and El-Beltagy, 2015).

In this work, we have adopted a supervised approach to determine the polarity of a sentence. Thus, to remedy the problem of the excessive size of the data representation vector, which is often of the lexicon size order, we have exploited the morphosyntactic tags of the lexicon words. As a result, instead of the information on the presence or the absence of the parsed sentence words in the lexicon, the vector contains only information related to word tags. Thus, the size of the vector becomes proportional to the number of tags. we also use the context of word negation to improve the accuracy of the analysis.

The remainder of the article is structured as follows. we begin by recalling in the second section some related works. Then, we present the linguistic resources in the third section. we give, in the next section, a description of the developed opinion detection approach. The last section will be devoted to the evaluation of the system.

2. Related work

This section presents previous work in the field of Arabic SA. It highlights the relevant characteristics, the main classification approaches and the best learning models.

The most used characteristics in SA are the occurrences of terms, the binary occurrences of terms, the term frequency (TF) and the term frequency–inverse document frequency (TFIDF) (Rahab *et al.*, 2018).

In (Aly and Atiya, 2013), the authors have used n-grams and TFIDF features to define the polarities of their LABR data source that contains reviews on Arabic books.

Similary, Mourad and Darwish combined several lexicons and used manually annotated tweets (Mourad and Darwish, 2013). They used bigrams and POS tags as features.

The work in (Ibrahim *et al.*, 2015) focuses on MSA and DA based on the following characteristics: syntactic features, negation and intensifiers.

The study presented in (Rushdi-Saleh *et al.*, 2011) performed tokenization and then used trigrams and TFIDF based on an Arabic corpus of opinions.

ElSahar and El-Beltagy have extracted a large lexicon from a multi-domain set of Arabic data (ElSahar and El-Beltagy, 2015). They based their study on different features among them the Delta-TFIDF.

Abdulla *et al.* tested a supervised approach that uses stems and a SVM classifier (Abdulla *et al.*, 2013). This allowed them to obtain better results than an unsupervised method based on the lexicon that was tested by the same system.

Finally, several studies have tested different classifiers and have shown that the classification with SVM classifiers gives better results compared to NB classifiers (Abuelenin *et al.*, 2018) or k-NN classifiers (ElSahar and El-Beltagy, 2015).

3. Resources

Concerning our approach, this section is devoted to present the used linguistic resources, the performed pre-processing and some statistics related to these resources.

3.1. Corpus

Our system exploits resources available in (ElSahar and El-Beltagy, 2015) to develop a classification system for the Arabic language. These resources consist of reviews belonging to several domains: hotels (HTL), products (PROD), movies (MOV) and restaurants (RES). They contain sentences labeled with three polarities: positive, negative and neutral and they also contain non-Arabic sentences.

As our system studies only the positive and negative polarities of Arabic reviews, the neutral polarity will not be taken into consideration. We have therefore filtered neutral and non-Arabic sentences.

Before using the corpus as input of our system, we normalized texts by deleting elongation (جميـــل to جميل), numbers, non-Arabic words and diacritics.

Table I shows the statistics of the corpus after filtering.

Domain	Number of positive and negative sentences	Number of neutral sentences	Number of non Arabic sentences
HTL	13408	2150	14
PROD	3595	308	369
MOV	1351	171	2
RES	10583	265	122

Table 1: Corpora statistics

3.2. Lexicon

The used lexicon was extracted by authors of (ElSahar and El-Beltagy, 2015) from the corpora mentioned above. The pre-processing step will adjust this lexicon to fit our system.

3.2.1. Lexicon pre-processing

we remove numbers, non-Arabic and dialectal words because our work focuses only on MSA. we also eliminate mutual redundancies between words, between words and bigrams and between words and trigrams. All pre-processing steps are automatic except dialectal words removal that is a manual task.

3.2.2. Lexicon statistics

The statistics on the lexicon after pre-processing are shown in Table 2. The lexicon contains an important rate of bigrams, some trigrams and its size differs from one domain to another. Moreover, in all domains the positive lexicon size is more important than the negative one.

Domain	Number of positive words	Number of negative words	Total
HTL	122	87	209
PROD	180	121	301
MOV	44	35	79
RES	282	241	523

Table 2: Lexicon statistics

4. Approach

In this section, we give a detailed description of the different steps of our system presented in Figure 1.

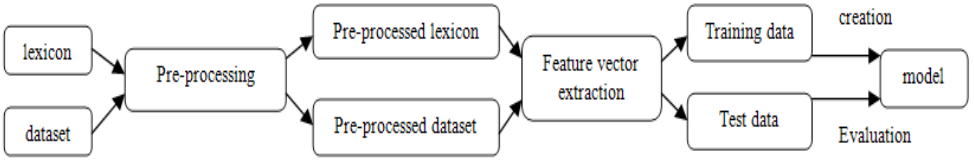

Figure 1: Classification process

4.1. pre-processing

In order to minimize the size of the data representation vector, which is often of the lexicon size order, the system starts by identifying the morphosyntactic tags of the lexicon and corpus words. Thus, we have used the POS Tagger developed in (Ababou and Mazroui, 2016) with a tag set limited to the verb and name tags.

Segmentation statistics are presented in Table 3, where V is the class of verbs and N is the class of nouns. The class V is divided into two subclasses: the first P_V contains the verbs corresponding to a positive feeling and the second N_V contains those corresponding to a negative feeling. Similarly, P_N is the subclass of the nouns corresponding to a positive feeling and the subclass N_N contains the nouns corresponding to a negative feeling.

Domain	Size of the class					
	P_V	P_N	of positive words	N_V	N_N	of negative words
HTL	21	101	122	11	76	87
PROD	46	134	180	36	85	121
MOV	10	34	44	3	32	35
RES	72	210	282	53	188	241

Table 3: Statistics of lexicon subclasses

4.2. Feature selection

Our system is based on three main features:

- *Vect={Occ_P_V, Occ_P_N, Occ_N_V, Occ_N_N}* is a vector consisting of the total number of occurrences in the analyzed sentence of the elements of each of the four classes P_V, P_N, N_V and N_N.
- *S_P*, *S_N* : represent respectively the sums of the positive and negative polarities of the analyzed sentence words. Thus *S_P = Occ_P_V + Occ_P_N* ,and *S_N = Occ_N_V + Occ_N_N* .
- *Nn*: consists in taking into consideration the presence of negation words. Thus, if a word is preceded by a negation word, then its polarity will be reversed. As a result, a positive word preceded by a negation word will be considered negative and vice versa. we used the following list of negation words:

  NW={لا»، «ولا»، «فلا»، «بلا»، «لست»، «فلست»، «ولست»، «وليست»،»لم»،»ولم»، «فلم»، «لن»،» «فلن»، «ولن»، «ليس»، «وليس»، «لات»، «ولات»، «فليست»، «فليس»، «غير»، «ليست»، «عديم»، «عدم»، «مضادة»، «دون»، «بدون»، «ليس له»، «عكس»، «ضد»، «اللا»، «مناف»، «غير قابل»، «خال من»، «لا يمكن»، «خالية من»}.

4.3. Model construction

Based on a combination of the three previous features, we generate an SVM model that we trained on the annotated corpus presented in the Section 3.1.2.

5. Evaluation

In this section, we evaluate our approach and compare its performance to that of a previous study conducted in (ElSahar and El-Beltagy, 2015). We first tested several models based on different combinations of the three features mentioned above to identify the most relevant feature set. The evaluation was performed on each of the following domains: HTL, PROD, MOV and RES. we constructed for each combination of features and for each domain an SVM supervised learning model from 80% of data extracted randomly from the annotated corpus associated with the domain. Then, we tested the obtained model on the remainder of data (20% of the annotated corpus). The used lexicons are also specific to the studied domains. For a given sentence, the main sets of the used features are:

Set 1 = {S_P, S_N} : we limit this set to the sums of the positive and negative polarities of sentence words.

Set 2 = {Occ_P_V, Occ_P_N, Occ_N_V, Occ_N_N} : we distinguish in this set the verbs from the nouns, and we calculate the sums of the polarities of each class. we notice that Set2 is a finer representation than Set1.

Set 3 =Set 1 $\cup$ *Nn* and *Set 4 =Set 2* $\cup$ *Nn* *:* in these sets, we take into consideration the context of negation which consists in changing the polarity of a word if it is preceded by a negation word. Table 4 presents the obtained test results using the accuracy as an evaluation metric for each domain and for each set of features.

Domain	Set1	Set2	Set3	Set4
HTL	89.37	90.71	91.27	91.83
PROD	83.31	82.61	86.23	86.23
MOV	80.74	80.74	81.85	82.22
RES	79.92	80.2	81.15	81.71

Table 4: Classification results using SVM

we notice that the use of word tags improves the performance of the system in most domains (comparing the results of columns Set1 and Set2). Moreover, the correction of word polarity by the use of negation words has also allowed to increase the accuracy in all domains (comparing the results of the column Set1 with those of the column Set3, and also the results of the column Set2 with those of the column Set4).

In order to illustrate the effect of the classifier nature on our system performance, we tested our approach with the NB classifier. We present in table V the obtained results with this classifier.

Domain	Set1	Set2	Set3	Set4
HTL	88.62	88.21	90.79	90.11
PROD	83.58	83.58	86.5	85.67
MOV	81.11	80.37	82.22	81.48
RES	79.49	79.82	81.05	80.49

Table 5: Classification results using NB

By comparing the results of Tables 4 and 5, we find that the results obtained with the two classifiers are close with a slight advantage for SVM. This confirms the conclusions of several studies that have shown the superior performance of the SVM classifier in the area of sentiment analysis.

The third experiment concerns the comparison of our system performances using the set4 features and the SVM classifier with those of the system developed in (ElSahar and El-Beltagy, 2015). we recall that the features used by this model are the count of words, the TFIDF and the delta TFIDF.

System	HTL	PROD	MOV	RES
Our system	91.83	86.23	82.22	81.71
ElSahar and El-Beltagy system	87.4	74.5	70.3	81.6

Table 6: Comparison between our system and system developed in (ElSahar and El-Beltagy, 2015)

Table 6 clearly illustrates the superiority of our system's performances compared with that developed in (ElSahar and El-Beltagy, 2015). Indeed, with the exception of the restaurant domain (RES) where the performances are close, the results of our system are much higher than those of (ElSahar and El-Beltagy, 2015).

6. Conclusion

In this article, we have developed a supervised approach to classify sentiments at the sentence level. It relies on various features such as the occurrence vectors of the sentence words present in the lexicon, the sum of polarities and the context of negation words.

Our target is to develop a lightweight approach in terms of feature vector size and execution time to classify data. This method is based on the distinction between lexical word nature using a POS Tagger.

Taking into consideration negation context also has a positive effect on the classification performance, because it helps to assign the correct polarity to each word.

The proposed method reveals a significant gain in term of accuracy, and surpasses the performance of a previous study.

w e plan in the future work to expand the studied classes by adding the neutral class and taking into account the intensifying words that intensify the feelings of the words that follow them, and the reducers that decrease those of the following words.

References

Ababou, N., Mazroui, A. (2016). A hybrid Arabic POS tagging for simple and compound morphosyntactic tags. *International Journal of Speech Technology*, doi: 10.1007/s10772-015-9302-8

Abdulla, N. A., Ahmed, N. A., Shehab, M. A., Al-Ayyoub M. (2013). Arabic Sentiment Analysis: Lexicon-based and Corpus-based. Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT´13). 6(12):1-6. doi: 10.1109/AEECT.2013.6716448.

Abdul-Mageed, M., Diab, M., Kübler, S. (2014). SAMAR: Subjectivity and sentiment analysis for Arabic social media. Computer Speech and Language. Elsevier Ltd. 28(1):20-37. doi: 10.1016/j.csl.2013.03.001.

Abdulla, N. A., Ahmed, N. A., Shehab, M. A., Al-Ayyoub M., Al-Kabi, M. N., Al-rifai, S. (2014). Towards Improving the Lexicon-Based Approach for Arabic Sentiment Analysis. *International Journal of Information Technology and Web Engineering*, 9(3):55-71. doi: 10.4018/ijitwe.2014070104.

Abuelenin, S., Elmougy, S., Naguib, E. (2018). Twitter Sentiment Analysis for Arabic Tweets. Proceedings of the International Conference on Advanced Intelligent Systems and Informatics. Vol. 639. doi: 10.1007/978-3-319-64861-3.

Aly, M., Atiya, A. (2013). LABR: A Large Scale Arabic Book Reviews Dataset. The 51st Annual Meeting of the Association for Computational Linguistics. pp. 494-498.

ElSahar, H., El-Beltagy, S. R. (2015). Building large arabic multi-domain resources for sentiment analysis. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 9042, pp. 23-34. doi: 10.1007/978-3-319-18117-2-2.

Ibrahim, H. S., Abdou, S. M., Gheith, M. (2015). Sentiment Analysis For Modern Standard Arabic And Colloquial. *International Journal on Natural Language Computing (IJNLC).* Vol. 4, No.2, April 2015, doi: 10.5121/ijnlc.2015.4207.

Mourad, A., Darwish, K. (2013). Subjectivity and Sentiment Analysis of Modern Standard Arabic and Arabic Microblogs. Proceedings of the 4th workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Vol. 3, pp. 55–64.

Rahab, H., Zitouni, A., Djoudi, M. (2018). SIAAC: Sentiment Polarity Identification on Arabic Algerian Newspaper Comments. Applied Computational Intelligence and Mathematical Methods. Vol. 662. doi: 10.1007/978-3-319-67621-0.

Rushdi-Saleh, M., Martín Valdivia, M. T., Ureña-López, L. A., Perea-Ortega, J. M. (2011). OCA: Opinion Corpus for Arabic. *Journal of the American society for information science and technology*. 62(10):2045-2054.

TICAM'18, p. 105 - 115

Approche non supervisée d'aide à une normalisation de l'écriture textuelle, basée sur un modèle augmenté Bi-LSTM et la prédiction des structures latentes : cas de l'amazighe & du tifinaghe

Hammou FADILI[1,2]
[1] Laboratoire CEDRIC du Conservatoire National des Arts et Métiers de Paris
192, rue Saint Martin, 75141, Paris cedex 3, France
[2] (Pôle Numérique, Programme Maghreb) de la FMSH Paris, 54, BD Raspail 75006, Paris, France
Hammou.fadili@cnam.fr fadili@msh-paris.fr

Résumé

Le traitement automatique de la langue amazighe (berbère) est compliqué à cause de ses nombreuses formes d'écritures, d'orthographes, de structures, etc., d'un côté, et le manque des données prétraitées et normalisées, de l'autre. Mettre en place des solutions pouvant aider à remédier à ces problèmes est un réel besoin et un grand défi pour le processus de normalisation que connait cette langue actuellement. C'est dans ce contexte que nous proposons une approche non supervisée basée sur l'apprentissage profond implémentant un système d'aide à la normalisation de l'écriture en général, par rapport à un contexte, des textes amazighes écrits en « tifinaghe ».

Mots clés : *Bi-RNN, Bi-LSTM, Attention, Point de vue, Normalisation, Ecriture, Thèmes latents, Structures latentes, Prédiction.*

1. Introduction

L'aménagement et la normalisation de la langue amazighe, en général, et de son écriture en tifinaghe, en particulier, sont une nécessité pour son développement. Cela concerne principalement les éléments de la composition linguistique : les structures syntaxiques, sémantiques, stylistiques et orthographiques. Dans un but de contribuer dans ce contexte, nous proposons une approche non supervisée basée sur l'apprentissage profond implémentant un système d'aide à la normalisation de l'écriture en général, par rapport à un contexte d'utilisation particulier, des textes « amazighes » écrits en « tifinaghe ».

L'approche est basée principalement sur l'amélioration du modèle des réseaux de neurones récurrents Bi-RNN (Bidirectional Recurrent Neural Networks) et son implémentation contextuelle à mémoires Bi-LSTM (Bidirectional Long Short-Term Memory) pour la prédiction des séquences. Nous y avons mis en place des mécanismes permettant de capter des signaux de l'écriture qu'on fait transiter et encoder à travers les différentes couches du

réseau pour l'optimisation de la prédiction des prochaines structures et les prochains mots complétant les phrases en cours de rédaction. Les structures latentes et l'orthographe de référence sont celles apprises à partir du corpus d'entrainement en tenant compte de leur importance et leur pertinence : intégration des notions d'attention et de point de vue dans les Bi-LSTM.

Le présent article est structuré comme suivant : dans la première partie, on y présente le modèle d'apprentissage, dans la deuxième partie, on y définit les éléments méthodologiques et technologiques implémentant l'approche. La troisième partie est consacrée à la génération du jeu de données pour l'apprentissage. Dans l'avant dernière partie on y présente les tests et les résultats obtenus.

2. Motivation

La normalisation de l'écriture d'une langue dont on parle ici concerne tous les aspects liés aux structures syntaxiques, sémantiques, stylistiques, et orthographiques. C'est une tâche complexe, notamment dans le cas des langues en voie de standardisation, telle que l'amazighe. Proposer des solutions pouvant aider les auteurs et les nouveaux utilisateurs de telles langues à respecter et à populariser leurs normes pourrait constituer un élément important dans leur processus d'aménagement et de standardisation.

Le retour et le développement rapides de l'Intelligence Artificielle rend de telles solutions possibles : on peut exploiter des approches non supervisées ne nécessitant pas de prérequis ni de données prétraitées, pour coder la science latente contenue dans le corpus étudié, comme source d'apprentissage et comme référence de la normalisation.

Ce sont ces éléments qui nous ont motivé à étudier et à proposer une approche non supervisée permettant de détecter et d'exploiter la « NORMALISATION » codée et contenue dans les textes de l'IRCAM : l'orthographe, la composition et les structures linguistiques. Notre apport consiste d'une part à adapter et à instancier le modèle contextuel des données (cf. Fadili, 2017) et d'autre part, à apporter des améliorations aux Bi-LSTM de base afin de contourner leurs limites dans la prise en charge des « métadonnées contextuelles ».

3. Le modèle d'apprentissage

Plusieurs études ont montré que, l'apprentissage profond a été exploité avec succès dans de nombreux domaines dont celui du traitement automatique de la langue naturelle (TALN). Une des meilleures implémentations est la génération d'espaces vectoriels sémantiques denses (Mikolov, 2013). D'autres réseaux tels que les réseaux de neurones récurrents (ou RNN pour Recurrent Neural Networks) ont également été améliorés et adaptés pour prendre en charge le caractère récurrent et séquentiel du traitement de la langue naturelle : chaque état est calculé à partir de son état précédent et de la nouvelle entrée. Ces réseaux ont l'avantage de propager l'information dans les deux sens : vers les couches d'entrées et vers les couches de sorties, reflétant ainsi une implémentation des réseaux de neurones, proche du fonctionnement du cerveau humain où l'information peut se propager dans tous les sens en exploitant le principe de la mémoire (cf. la version LSTM des RNN dans ce qui suit), via des connections récurrentes propageant l'information d'un apprentissage ultérieur (l'information

mémorisée). Ce sont ces caractéristiques qui leur permettent une meilleure prise en charge de plusieurs aspects importants de la langue naturelle. En effet, ils ont cette capacité de capituler les structures latentes syntaxiques, sémantiques, stylistiques et orthographiques, à partir de l'ordre des mots, contrairement aux autres technologies telles que celles basées sur la notion de sacs de mots où aucun ordre n'est considéré, impliquant bien évidemment la perte d'information associée.

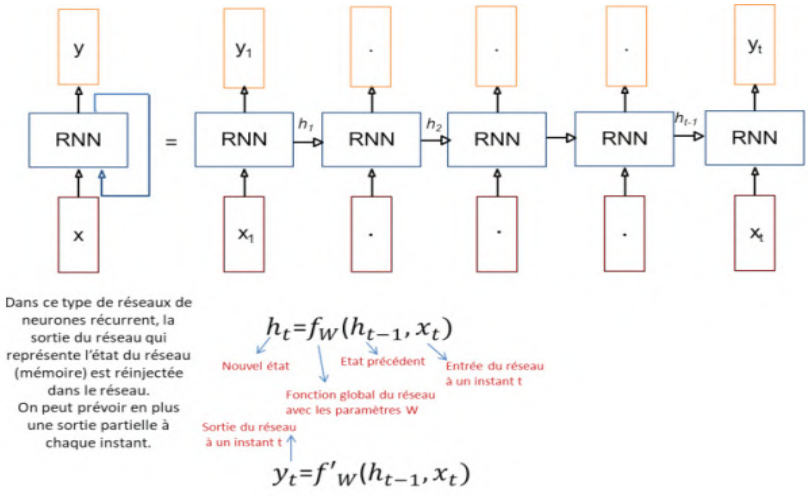

Figure 1: Fonctionnement d'un RNN

Dans les RNN en série, chaque nouvel état interne et chaque nouvelle sortie dépendent simplement de la nouvelle entrée et de l'ancien état.

Les RNN peuvent aussi être empilés et bidirectionnels, et les équations simples précédentes peuvent être redéfinies pour les 2 sens d'apprentissage suivant le modèle ci-après.

$$\textit{Vers l'avant}: \; h_t^l = \tanh W_t^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}, h_t^l \in \mathbb{R}^n, W^l \textit{ matrice de dimension } [n \times 2n]$$

$$\textit{Vers l'arriere}: \; h'^l_t = \tanh W'^l_t \begin{pmatrix} h_t^{l-1} \\ h'^l_{t+1} \end{pmatrix}, h_t^l, h'^l_t \in \mathbb{R}^n, W'^l \textit{ matrice de dimension } [n \times 2n]$$

$$\textit{Sortie}: \; y_t = \tanh W_t \begin{pmatrix} h_t^l \\ h'^l_{t+1} \end{pmatrix}, y_t \in \mathbb{R}^n, W^l \textit{ matrice de dimension } [n \times 2n]$$

L'entrainement des RNN de plusieurs couches se fait, comme pour les autres types de réseaux, par la minimisation de l'erreur (différence entre la sortie désirée et la sortie obtenue) qu'on obtient par la rétro-propagation de l'erreur et la descente du gradient. On peut démontrer mathématiquement que la profondeur des RNN pouvant être élevée, à cause de leur nature séquentielle, dépendant en général du nombre de mots à traiter à la fois ; peut provoquer :

- Soit l'évanouissement du gradient (Vanishement of Gradient) dans les premières couches et l'arrêt de l'apprentissage à partir d'une certaine profondeur. Dans le cas où l'on doit multiplier le gradient, un nombre élevé de fois, par un poids w / $|w|<1$.

- Soit l'explosion du gradient (Explosion of Gradient) toujours dans les premières couches et l'arrêt de l'apprentissage à partir d'une certaine profondeur. Dans le cas où l'on doit multiplier le gradient, un nombre élevé de fois, par un poids w / $|w|>1$.

L'architecture LSTM permet de remédier à ces problèmes (Hochreiter, & Schmidhuber, 1997). Elle est basée sur un contrôle plus fin du flux d'information dans le réseau, grâce à trois portes : la porte d'oubli qui décide de ce qu'on doit effacer de l'état (h_{t-1}, x_t), la porte d'entrée qui choisit ce qu'on doit additionner à l'état et la porte de sortie qui choisit ce qu'on doit garder de l'état (cf. équations ci-dessous).

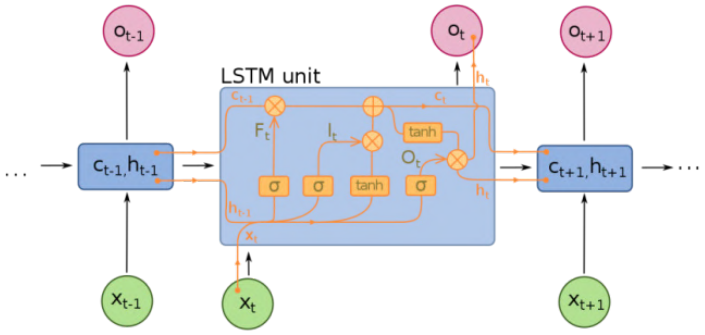

Figure 2: Fonctionnement d'un LSTM de base

$$
\begin{aligned}
F_t &= \sigma(W_F x_t + U_F h_{t-1} + b_F) && \text{(forget gate)}\\
I_t &= \sigma(W_I x_t + U_I h_{t-1} + b_I) && \text{(input gate)}\\
O_t &= \sigma(W_O x_t + U_O h_{t-1} + b_O) && \text{(output gate)}\\
c_t &= F_t \circ c_{t-1} + I_t \circ \tanh(W_c x_t + U_c h_{t-1} + b_c)\\
h_t &= O_t \circ \tanh(c_t)\\
o_t &= f(W_o h_t + b_o)
\end{aligned}
$$

Ces équations définissant le processus d'apprentissage d'un LSTM expriment le fait que ce type de réseaux permet d'annuler certaines informations inutiles et d'en renforcer d'autres ayant un grand impact sur les résultats. On peut montrer également par des calculs mathématiques que cette architecture, permet, en plus de l'optimisation des calculs dans le réseau, de résoudre les problèmes liés à l'évanouissement et à l'explosion du gradient. C'est ce qui a motivé notre choix à utiliser et à améliorer ce modèle en l'adaptant à nos besoins. Nous lui y avons intégré en plus la notion de perspective ou de point de vue d'analyse ainsi que la notion de l'attention dans le processus général de la prédiction.

Notre modèle permet en plus de contrôler le flux dans le « Contexte » i.e. :

- Ce qu'il faut oublier de l'état
- Ce qu'il faut utiliser de l'état
- Ce qu'il faut envoyer en sortie
- Suivant un point de vue ou la perspective
- En portant attention aux sorties pertinentes

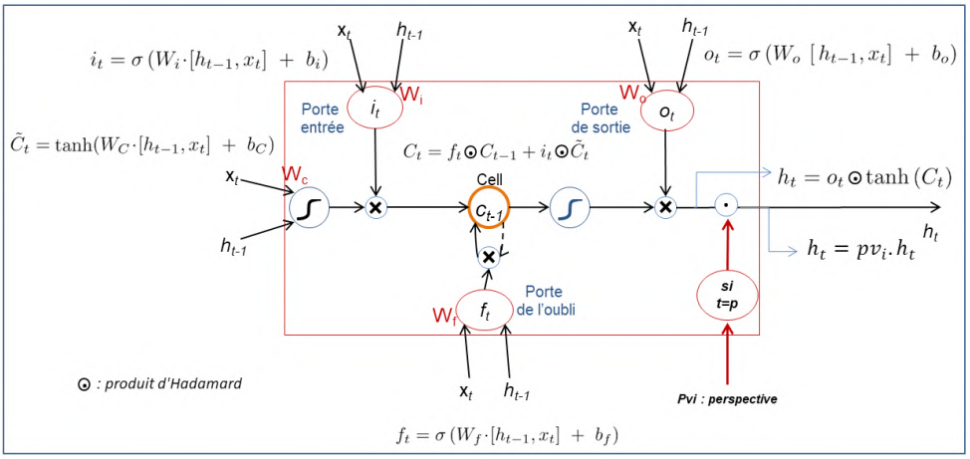

Figure 3: Architecture d'un LSTM améliorée

Cette dernière notion est représentée en dehors de l'architecture interne du modèle LSTM (cf. l'architecture générale de l'approche).

4. L'intégration des éléments technologiques

Notre approche a pour objectif de prédire et de suggérer de compléter les phrases en cours de rédaction par les mots les plus pertinents dans leurs formes orthographiques normalisées et suivant une structure linguistique bien particulière (celle contenue dans les textes du corpus). Autrement dit, à une séquence de mots que l'utilisateur est entrain de saisir, le système lui propose une séquence de mots, des phrases ou des bouts de phrases normalisés. Cela exige l'utilisation de la version séquence-à-séquence (sequence-to-sequence ou seq2seq) des Bi-LSTMs. L'architecture proposée est constituée des couches principales suivantes :

- Encodage
  - Prétraitement
  - Représentation interne
  - Domaine (Point de vue)
- Décodage
  - Nouvelle représentation interne (calculée)
  - Attention
  - Prédiction.

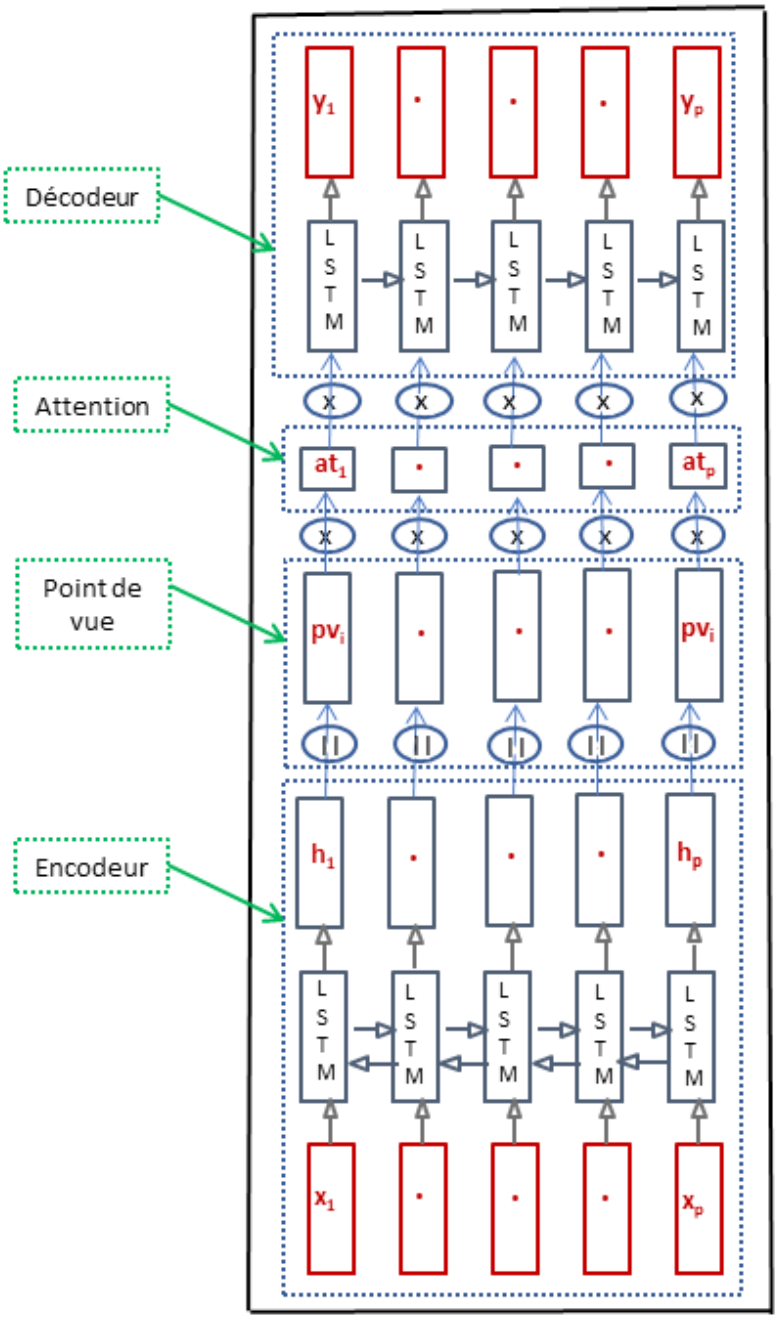

Figure 4: Architecture

Une perspective est un ensemble de mots caractérisant un point de vue d'analyse (espace multidimensionnel).
Soit *p* le nombre de mots de la phrase courante
Soit *i* l'indice de perspective courante
Soit *P* une perspective définie par les dimensions $p_1, p_2, ..., p_i$
Soit *H* la matrice composée par les représentations cachées $h_1, h_2, ..., h_p$ générées par le LSTM

Considérons $pv_i = \sum_1^p d_j \text{ où } d_j = \sum_1^l d_k \text{ et } d_k = \frac{p_k . h_j}{\| \ldots \| \| \ldots \|}$

pv_i est une première pondération dans le calcul de la prédiction. Elle est basée sur la somme des distances *cosinus* entre les représentations cachées générées par le LSTM et les dimensions de la perspective d'analyse ciblée.

Cela permet d'obtenir une première transformation, pondérée par le rapprochement à la vue considérée, des représentations internes des sens de la phrase :

$$H' = \left(h_1', h_2', \ldots, h_p'\right), \text{où } h_j' = d_j . h_j$$

d_j *est la somme des ditances cosinus de hi à chaque dimension.*

$$\propto_d = softmax\left(\sum at_i\right)$$

$at_i = w_i^T . h_i'$ *constitue l'attention apprise par* (w_i) *pour chaque* h_i'

$prediction = H \propto_d^T$ *représente la sortie du système.*

5. Corpus et jeu de données

La constitution du corpus s'est faite par la collecte d'un nombre important de documents écrits en tifinaghe, obtenus principalement de la version amazighe du site de l'IRCAM (plusieurs dizaines de pages). L'obtention du modèle des données d'apprentissage s'est faite à partir d'une version simplifiée du modèle de langue et le modèle sémantique étendu (Fadili, 2017) et des projections des représentations vectorielles initiales des mots relatives à un espace de grandes dimensions (taille du vocabulaire), dans un espace sémantique de dimensions réduites en utilisant la technologie word2vec (w2v). Le but est de créer un modèle d'instances enrichi et adapté au contexte de la normalisation de la langue, doté d'une représentation vectorielle de taille raisonnable, indispensable pour l'optimisation des calculs. L'instanciation du modèle s'est faite par le découpage des textes en phrases, puis la génération pour chaque mot des n-grams de ses fenêtres contextuelles :

- 2-grams
- 3-grams
- 4-grams
- 5-grams.

Cette première version du jeu de données est ensuite enrichie par d'autres paramètres pour prendre en charge la structure syntaxique et la distribution thématique. Nous avons associé à chaque mot du jeu de données sa catégorie grammaticale (Part Of Speech POS) (Outahajala *et al.*, 2011) ainsi que sa distribution thématique (Topics) obtenue par LDA (Latent Dirichlet Allocation) (Bei *et al.*, 2003). Les autres structures latentes sont prises en charge par le codage du caractère séquentiel des textes par les LSTMs. L'orthographe est également codée à partir du corpus d'entrainement. C'est ce jeu de données qui est fourni à notre modèle Bi-LSTM augmenté. Ce modèle a l'avantage de tenir compte, en plus, du contexte local (n-grams) et du contexte global (thèmes), des mémoires long-court-termes et d'ordre des mots qui encode les structures latentes des textes (syntaxique, sémantique, etc.).

6. Expériences & évaluations

Nous avons développé un certain nombre de modules et de programmes informatiques, implémentant les éléments du processus général, dont principalement :

- L'extraction automatique des différentes caractéristiques (Features) (cf. jeu de données).
- Implémentation d'un système d'apprentissage profond de type Bi- LSTM doté des mécanismes d'attention et de domaine, décrits précédemment.

Nous avons également conçu un environnement informatique centralisant l'accès à tous les modules :

- Intégration de tous les éléments dans un seule Workflow implémentant tous les modules des traitements.

Extraction des caractéristiques :

Nous avons développé et mis en place 3 modules qui tournent dans le même environnement « python-Canvas » et exploité un processus basé sur une « Méta-Jointure générale » du même environnement pour faire le lien entre les différents composants.

Sans rentrer dans les détails, les 3 modules consistent, dans l'ordre, à extraire les caractéristiques suivantes :

- Le premier permet d'extraire le « contexte linguistique » de chaque mot à travers une fenêtre glissante de taille paramétrée n.
- Le deuxième permet d'annoter grammaticalement tous les mots du texte : effectuer le POS Tagging (Part Of Speech Tagging)
- Le troisième implémente la technologie « Topic modelling et LDA » afin d'extraire automatiquement le ou les domaine(s) traité(s).

Workflow :

L'implémentation du Workflow a été effectuée dans la plateforme Orange Canvas. Elle automatise l'enchainement des résultats des modules précédents pour l'extraction des caractéristiques et la génération du modèle d'instances prêt à être exploité pour l'apprentissage.

Réseau de neurones Bi-LSTM amélioré :

Nous avons développé à partir de zéro, un réseau de neurones de type bi-LSTM personnalisé et adapté à nos besoins. Il s'agit d'un réseau de neurones (multicouches & récurrent de type Bi-LSTM doté de mécanismes d'attention et de domaine).

Afin d'assurer l'honnêteté du système, nous avons séparé les données d'apprentissage générées en trois parties :

Une première partie pour la validation sur 20% du jeu de données, afin d'optimiser les hyper-paramètres du système : le pas d'apprentissage, le type de la fonction d'activation et le nombre de couches.

Le reste du jeu de données est divisé en deux parties :

- 60% pour l'entrainement, afin d'estimer les meilleurs coefficients (wi) de la fonction du réseau de neurones, minimisant l'erreur entre les sorties réelles et les sorties désirées.
- 20% pour les tests, afin d'évaluer les performances du système.

Pendant toutes les phases d'apprentissage, le système est autonome, il est capable de générer les caractéristiques du texte (Features) pour l'entrainement.

Résultats

=== Sample of Dataset===

«Column1»,»Column2»,»Column3»,»Column4»,»Column5»,»Column6»,»Column7»,»Column8»,»Column9»
«691»,»1»,»1»,»0.250»,»ⴹⵣⵙⵔⴰ»,»ⴹⵣⵎⵏ»,»ⴹ»,»ⴹⵔⵔⵔⵙⵠⴰⵔ»,»»
«692»,»1»,»1»,»0.250»,»ⴹⵣⵙⵔⴰ»,»ⴹⵣⵎⵏ»,»ⴹ»,»ⴹⵔⵔⵔⵙⵠⴰⵔ»,»ⴹⵔⴹⴰⵠⵙⵊⴹⵏ»
«695»,»1»,»1»,»0.250»,»ⴹⵣⴻⴰⵇⵏ»,»ⴰⵠ»,»ⵔⵔⵠⵔⵛⵙⵏ»,»ⴰⵠⵢⴰⵔ»,»»
«696»,»1»,»1»,»0.250»,»ⴹⵣⴻⴰⵇⵏ»,»ⴰⵠ»,»ⵔⵔⵠⵔⵛⵙⵏ»,»ⴰⵠⵢⴰⵔ»,»ⵏ»
«699»,»1»,»1»,»0.125»,»ⴹⵣⵎⵏ»,»ⴹ»,»ⴹⵔⵔⵔⵙⵠⴰⵔ»,»ⴰⵔⴹⴰⵠⵙⵊⴹⵏ»,»ⵖⵙⵔⵊⵏ»
«701»,»1»,»1»,»0.125»,»ⴹⵣⵎⵏ»,»ⴹ»,»+ⴹⵢⴰⵍⵔⵔⵍⵊⵏ»,»+ⴹⵏⵠⵢⴰⵔⵔⵏ»,»»
«702»,»1»,»1»,»0.125»,»ⴹⵣⵎⵏ»,»ⴹ»,»+ⴹⵢⴰⵍⵔⵔⵍⵊⵏ»,»+ⴹⵏⵠⵢⴰⵔⵔⵏ»,»ⵏⴹⵢ»
«706»,»1»,»1»,»0.250»,»ⴹⵣⵔⵣⵎⴰⵖⵏ»,»ⵏⵔ»,»ⴹⵢⵔⵔⴰⵏ»,»+ⴰⵣⵔⵍⵔⵔ+»,»ⵏ»

=== Sample of Predictions on test split ===

inst#,actual,predicted,error,prediction
8,24:ⵠ,24:ⵠ,1
9,170:+ⵔⴰⵣⵢ+,170:+ⵔⴰⵣⵢ+
10,145:+ⴰⵔⵔⵔⴰⵠⵙⵔ+,145:+ⴰⵔⵔⵔⴰⵠⵙⵔ+
11,53:ⴹⵔⵠⵠⴰⵖⵏ,53:ⴹⵔⵠⵠⴰⵖⵏ
12,16:ⴰⵔⵔⵙ++ⵠ,16:ⴰⵔⵔⵙ++ⵠ
13,77:ⴹⵜ+,77:ⴹⵜ+
14,79:ⴹ++ⵙⵣⵎⴰⵖⵏ,94:!ⴹⵢ
15,33:ⵅⵃ,33:ⵅⵃ,
16,93:!,93:!
17,23:ⵠ,23:ⵠ
18,4:ⴰⵠ,4:ⴰⵠ

=== Evaluation on test split ===

Time taken to test model on test split: 0.75 seconds

=== Summary ===

Correctly Classified Instances	455	83.7937 %
Incorrectly Classified Instances	88	16.2063 %
Kappa statistic	0.8323	
Mean absolute error	0.0021	
Root mean squared error	0.0302	
Relative absolute error		21.9222 %
Root relative squared error		44.2078 %
Total Number of Instances	543	

=== Sample Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1,000	0,002	0,500	1,000	0,667	0,706	0,999	0,500	ⵉⴰⵉⴰⵓⵅⵉⵅⵉ
0,600	0,000	1,000	0,600	0,750	0,773	0,799	0,604	ⵉⴷⵉⵍ
1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	ⵉⵅⵙⵏ
0,500	0,000	1,000	0,500	0,667	0,706	0,750	0,502	ⵉⵍⵍₒ
1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	ⵉⵍⵛⴰₒⴰ
1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	ⵉⵍⵣⵍ
1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	ⵉⵛₒⵉⵍₒⵍⵍ
0,500	0,000	1,000	0,500	0,667	0,706	0,750	0,502	ⵉⵛₒⵙₒ
1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	ⵉⵛₒⵙⵙⵏ

Après l'entrainement du système sur le corpus constitué, nous avons effectué des tests sur un petit texte constitué de 543 mots. Le système a réussi à prédire correctement l'emplacement de 455 mots dans les différentes phrases du texte, en tenant compte des structures latentes codées à partir des textes. Le système a par contre échoué sur les 88 mots restants.

Pour plus d'information sur la lecture de ces résultats cf. (lire w eka) dans les références.

7. Conclusion

Les expériences menées confirment l'apport, dans de bonnes proportions, de l'approche proposée, pour prédire et aider à la normalisation de l'écriture et de l'orthographe de l'amazighe en tifinaghes.

Comme décrit précédemment, nous avons ajusté plusieurs hyper-paramètres pour optimiser le système et générer tous les paramètres des instances. Sur la totalité du corpus obtenu, nous avons validé l'approche sur 20% du corpus, entrainé le système sur environ 60% et testé sur environ 20%. Les différentes mesures montrent les bonnes performances de l'approche, en termes de précision et de performance.

Cela, grâce principalement aux améliorations apportées à l'intégration de certains aspects de la langue et à des extensions mises en place pour le Bi-LSTM :

- Modèle de langue et son instanciation
- Introduction des notions d'attention, de domaine et de point de vue dans les Bi-LSTM.

Références

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, pp. 993-1022.

Demšar, J., Zupan, B., Leban, G., Curk, T. (2004). Orange: From experimental machine learning to interactive data mining. In European Conference on Principles of Data Mining and Knowledge Discovery, pp. 537-539. Springer, Berlin, Heidelberg.

Fadili, H. (2017). Retour d'expérience : l'utilisation de l'apprentissage profond (deep learning) dans le contexte de l'analyse sémantique des langues peu dotées (DiLiTAL 2017).

Fang, D., Yang, H., Gao, B., Li, X. (2018). Discovering research topics from library electronic references using latent Dirichlet allocation. Library Hi Tech.

Harris, C. (2018). Searching for Diverse Perspectives in News Articles: Using an LSTM Network to Classify Sentiment.

Hochreiter, S., &Schmidhuber, J. (1997). LSTM can solve hard long time lag problems. In Advances in neural information processing systems, pp. 473-479.

Outahajala M, Benajiba Y, Rosso P, Zenkouar L (2011). Pos tagging in Amazighe using support vector machines and conditional random fields. In: Natural Language Processing and Information Systems, Springer Berlin Heidelberg, pp. 238–241

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pp. 3111-3119.

Sutskever, I., Vinyals, O., Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pp. 3104-3112.

Lire w eka, http://www.plantie.fr/EMA/projetdatamining/fouille2010-1/docs/lirew eka.pdf

TICAM'18, p. 117 - 125

Modèle Word2vec pour la langue amazighe

Mohamed BINIZ[1], Samir BOUKIL[2],
Rachid EL AYACHI[1] , Mohamed FAKIR[1]
[1] Laboratoire de traitement de l'information et aide à la décision
Faculté des Sciences et Techniques, Université Sultan Moulay Slimane, Béni Mellal, Maroc
[2] Laboratoire de recherche en optimisation, systemes emergents, reseaux et imagerie
Université Chouaïb Doukkali, El Jadida, Maroc
mohamedbiniz@gmail.com
{boukilsamir, fakfad}@yahoo.fr
rachid.elayachi@usms.ma

Résumé

Les méthodes d'apprentissage dans le domaine de traitement automatique des langues naturelles se maintiennent de plus en plus sur des représentations vectorielles des mots (Bag of word, TF-IDF, … etc.). Ces techniques, déjà employées avec succès dans de nombreuses tâches pour la langue amazighe sont capables de représenter des mots ainsi que les relations les liant. De manière générale, ces méthodes utilisent des représentations qui traitent tous les mots d'un contexte de façon égale. Ce travail propose une méthode qui s'appuie sur les modèles de contextes continus en intégrant la position relative des mots dans un contexte et les techniques de Big Data pour construire un modèle word2vec pour l'amazighe. Les résultats montrent que l'utilisation des techniques Big Data permet de gagner jusqu'à 91 % de temps d'exécution pour construire un modèle word2vec amazigh à partir d'un corpus extrait de l'internet.

Mots clés : w ord2vec, TAL, Amazigh.

1. Introduction

Dans le domaine de traitement automatique de la langue naturelle (la classification des documents (Boukil *et al.*, 2017 ; Samir *et al.*, 2018), l'analyse des sentiments (Alayba *et al.*, 2018 ; Singh *et al.*, 2014), l'indexation sémantique (Youness *et al.*, 2018). la récapitulation des documents (Al Qassem, 2017)), on trouve plusieurs manières pour représenter et coder le texte afin d'extraire les descripteurs les plus pertinents au sens du problème à résoudre. Les algorithmes d'apprentissage ne sont pas capables de traiter directement les textes, plus généralement, les données non structurées comme les images, les sons et les séquences vidéo. C'est pourquoi, une étape préliminaire dite de représentation est nécessaire. Cette étape consiste généralement en la représentation de chaque document par un vecteur, dont les composantes sont par exemple les mots contenus dans le texte, afin de le rendre exploitable par les algorithmes d'apprentissage. Une collection de textes peut être ainsi représentée par une matrice dont les lignes sont les termes qui apparaissent au moins une fois et les colonnes sont les documents de cette collection.

Un grand nombre de chercheurs dans le domaine ont choisi d'utiliser une représentation vectorielle « sac de mot » dans laquelle chaque texte est représenté par un vecteur de n mots dont sa valeur est le nombre de son apparition dans un document. Ces comptages de mots nous permettent de comparer des documents et d'évaluer leurs similitudes pour des applications telles que la recherche, la classification de documents et la modélisation de sujets.

D'autres chercheurs utilisent la fréquence de document inverse-fréquence de terme (TF-IDF) qui est une autre façon de représenter les mots d'un texte. Avec TF-IDF, les mots sont pondérés, elle mesure la pertinence, pas la fréquence. En d'autres termes, les compteurs de mots sont remplacés par des scores TF-IDF sur l'ensemble des données.

L'utilisation de ces techniques invoque le problème de la taille du vocabulaire, car en utiliseant tous les mots présents dans les documents de l'espace d'apprentissage, on se retrouve face à un espace vectoriel ayant une dimension très large. Le traitement d'un tel espace nécessitera beaucoup de mémoire, de temps de calcul et pourra empêcher l'utilisation des algorithmes d'apprentissage plus puissants. Pour résoudre ce problème, on utilise le modèle word2vec pour réduire la taille du vocabulaire qui a pour objectif de sélectionner ou d'extraire un sous-ensemble optimal de caractéristiques pertinentes.

2. État de l'art

2.1. Sac de mots

Les textes sont transformés simplement en vecteurs dont chaque composante représente un mot (Mahdi *et al.*, 2014). Utiliser les mots comme termes a comme avantage d'exclure toute analyse grammaticale et toute notion de distance entre les mots, mais présente plusieurs inconvénients qui sont les suivants :

Cette représentation ne respecte pas la sémantique du mot. Par exemple: les mots «vélos» et «Bicyclette » sont souvent utilisés dans le même contexte. Cependant, les vecteurs correspondant à ces mots sont orthogonaux dans le sac de mots modèle.

Le problème devient plus sérieux lors de la modélisation des phrases. Ex: « Buy used bikes» et « Buy old bicycles ». Elles sont représentés par des vecteurs orthogonaux dans le modèle Bag-of-words.

2.2. TF-IDF

La pondération des termes permet de mesurer l'importance d'un terme dans un document. Cette importance est souvent calculée à partir de considérations et interprétations statistiques. L'objectif est de trouver les termes qui représentent mieux le contenu d'un document. Pour calculer la pondération, on distingue les méthodes suivantes :

2.2.1. Mesure TF (Term Frequency)

Cette mesure est proportionnelle à la fréquence du terme dans le document. Elle peut être utilisée telle quelle ou selon plusieurs déclinaisons (log(tf), présence/absence,. . .).

2.2.2. Mesure TF-IDF (Term Frequency Inverse Document Frequency)

- idf (Inverse of Document Frequency) : idf = $\log(^{N}/_{Df})$

Où:

- Df: Le nombre de documents contenant le terme.
- N : Le nombre total de documents de la base documentaire.

Un terme qui apparaît souvent dans la base documentaire ne doit pas avoir le même impact qu'un terme moins fréquent. La mesure TF-IDF est une bonne approximation de l'importance du terme dans le document, particulièrement dans les corpus de documents de taille homogène. La mesure TF-IDF est calculée comme suit :

- TF-IDF est basé sur le modèle de sac de mots (Bow), donc il ne capture pas la position dans le texte, les cooccurrences dans différentes phrases, etc.
- Pour cette raison, TF-IDF n'est utile que comme fonctionnalité de niveau lexical
- Impossible de capturer la sémantique.

2.3. modèle Word2vec

Les représentations vectorielles w ord2Vec sont construites automatiquement à partir de ressources textuelles volumineuses constituées de textes bien formés à partir de l'étape de prétraitement. L'objectif de ces représentations est de prédire le mieux possible le contexte des mots. En conséquence, la similarité entre les représentations w ord2Vec de deux mots intègre à la fois une similarité sémantique et une similarité de rôle dans la phrase. Ce modèle w ord2Vec a deux architectures : le sac continu de mots (CBOw) et Skip-gram (Mikolov *et al.*, 2013).

2.4 Modèle CBoW

Le modèle du CBOw (Rong, 2014) n'est rien de plus qu'un réseau neuronal linéaire formé pour prédire un mot à partir du contexte dans lequel il se produit. La couche d'entrée de ce réseau de neurones utilise des "sacs de mots" binaires représentant une fenêtre de contexte. Dans cette configuration, les vecteurs d'entrée sont des vecteurs de la taille du vocabulaire avec un 1 dans la colonne i si le mot i est présent dans le document, 0 sinon. Chaque mot est alors projeté dans la matrice globale, l'ensemble des représentations est ensuite additionné pour former une unique couche cachée figure 1. Cette couche cachée passe par la couche de sortie et les fonctions d'activation type "Softmax" (Rong, 2014) tentent de prédire le mot au cœur de la fenêtre. L'erreur de prédiction est ensuite utilisée pour corriger les matrices de poids via une rétropropagation de gradient.

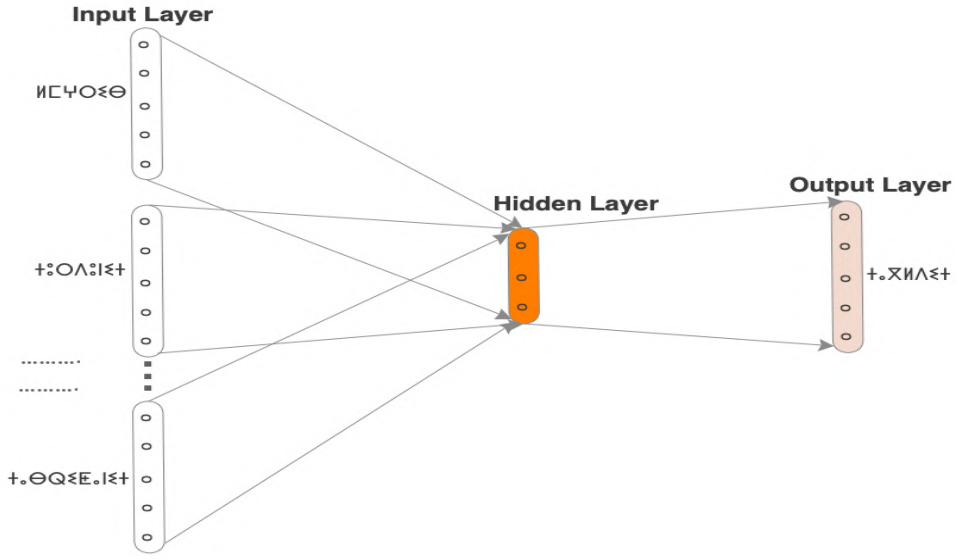

Figure 1: Représentation d'un modèle CBoW

2.5. Modèle Skip-gram

Le modèle Skip-gram (Rong, 2014) est également un réseau de neurones artificiels simple pour prédire les « mots voisins » d'un mot spécifique au milieu d'une phrase, contrairement au CBOw. Les probabilités de sortie vont se rapporter à la probabilité de trouver chaque mot vocabulaire dans notre mot d'entrée.

Par exemple, si nous avons donné le mot d'entrée «ⵜⴰⵖⵍⴰⵣⵜ» au réseau formé, les probabilités de sortie seront beaucoup plus élevées pour les mots « ⵍⵎⵢⵓⵣⵏ» et « ⵜⴰⵛⵓⴰⵏⵜ « que pour les mots sans rapport «ⵜⵗⵙ» et «ⴰⵢⴰⵓⴰ».

La couche d'entrée du réseau ne contient donc que la représentation en sac de mots binaire du mot au cœur du contexte («ⵜⴰⵖⵍⴰⵣⵜ «). Ce mot est projeté dans la matrice de poids globale, puis transmis à la couche de sortie qui va prédire un mot. Cette prédiction est ensuite corrigée par rétro- propagation pour chacun des mots de la fenêtre de contexte.

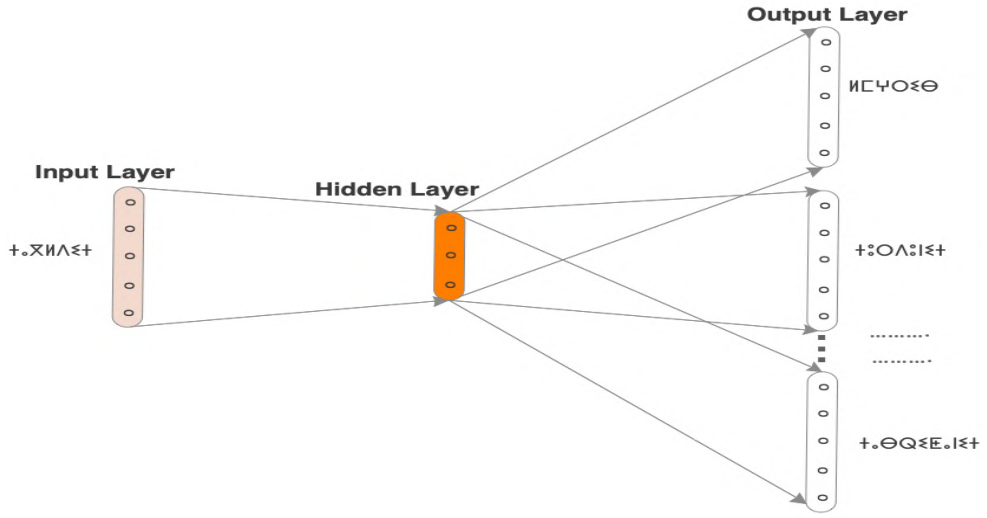

Figure 2: Représentation d'un modèle Skip-Gram

3. **Expérimentations**

Nous avons formé deux modèles : CBow et Skip-gram. Ce modèle a appris à représenter les mots sous une forme numérique, cette représentation étant un vecteur de 300 dimensions. Pesez ce vecteur comme les coordonnées de la représentation de chaque mot dans l'espace. Nous pouvons alors analyser comment ces représentations sont disposées dans cet espace. Notre espoir est que les mots ayant une charge sémantique et syntaxique similaire soient proches les uns des autres.

3.1. Corpus

Le corpus est une collection de textes d'amazigh écrite en Tifinagh, qui couvre la langue moderne d'amazigh utilisée dans les articles de journaux. Le texte contient des mots alphabétiques, numériques et symboliques.

Le corpus est constitué de 2038 documents structurés en fichiers texte et collectés à partir de 1 journal en ligne arabe: Agence Marocaine de Presse. Il a fallu plus de dix heures pour traiter tout ce texte en utilisant une seule machine. Pour cette raison nous avons invoqué les calculs distribués et les technologies Big Data afin de réduire le temps de calcul.

3.2. Architecture du système proposé

L'architecture du système proposé est illustrée par la figure 3.

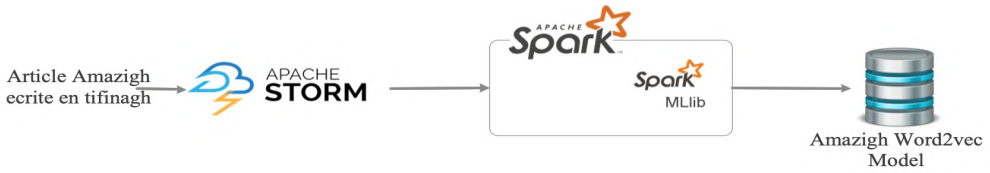

Figure 3: Architecture de système proposé

1. Téléchargement des données des articles depuis l'internet.
2. Démarrage d'Apache Storm et la lecture des fichiers de données téléchargé. Cet étape contient un ensemble de sous-étapes :
   - Au début, nous avons besoin d'un bec qui reçoit des données du flux de texte et apporte les données dans notre topologie, nommé amazighDumpSpout.
   - L'amazighDumpSpout envoie les tuples de données à un boulon appelé BoltPlainTextExtraction qui supprime toutes les étiquettes tags et de formater en texte le flux de données.
   - Dans la dernière étape, les données doivent être envoyées à BoltSentenceExtraction qui extrait les phrases depuis le texte dans les articles et convertir les résultats de l'analyse morphologique en séquences. Donc, notre topologie est schématisée par la figure 4 suivante :

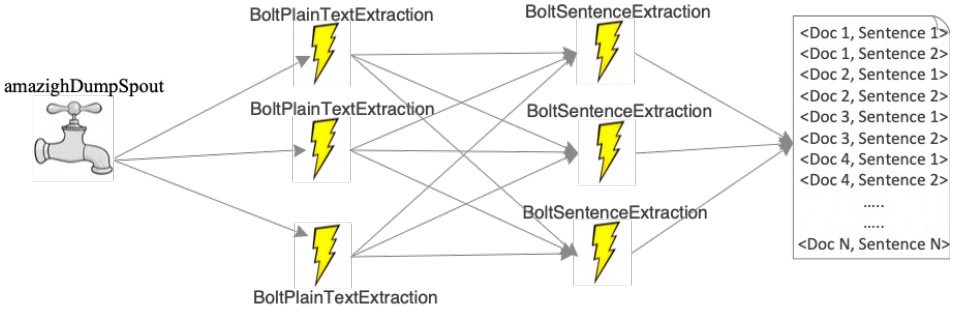

Figure 4: Topologie d'apache storm

Transmettez un tableau à l'implémentation w ord2Vec de Apache Spark Mllib (Karim *et al.*, 2017 ; Luu, 2018) et faites-lui apprendre le modèle.

Les étapes 2 à 3 sont des parties à exécuter sur Apache Storm (Bhatnagar et Hart, 2015 ; Jain, 2017) et Apache Spark. Ce qui nous permet de raccourcir considérablement le temps de traitement de la phase d'apprentissage en activant le traitement parallèle.

D'après la figure 5, nous observons que le système a besoin de 10 heures en utilisant un seul nœud. Pourtant lorsqu'en utilisant 6 nœuds, le système a besoin seulement d'une 50 minutes.

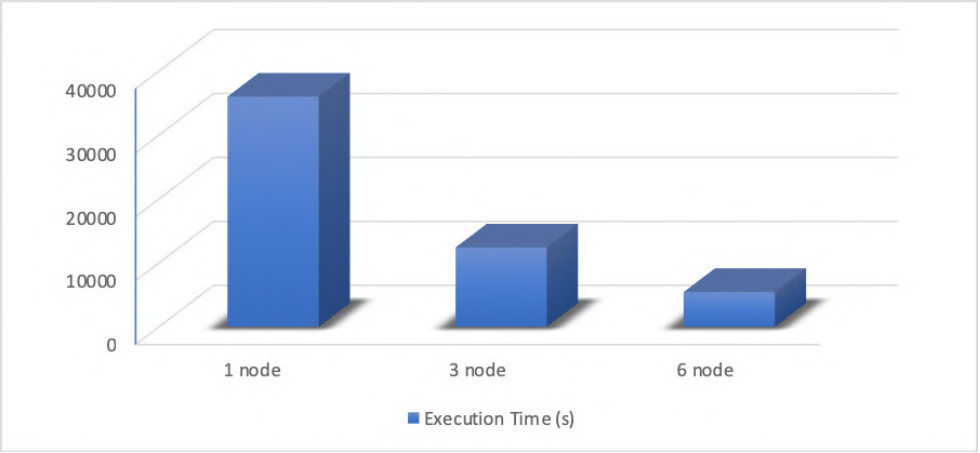

Figure 5: Temps d'exécution en seconde (s)

Pour le test de modèle élaboré, on commence par l'identification des voisins les plus proches de certains mots dans l'espace des représentations.

ⵜⴰⵎⵢⴰⵡⵜ	ⵣⵓⵔⵣⴻ	ⵜⴰⵍⵜⴰⵙⵜ
ⵜⴰⵎⴻⴻⵙⴻⵜ ⵜⴰⵀⵓⵙⵅⵜ ⵜⴰⵎⵢⴰⵡⵜ ⵜⴰⵀⵓⵙⵅⵜ	ⵣⵎⵍⵣⵜⵍ ⴰⵎⵍⵣ ⵣⵓⵔⵣⴻ ⵙⵎⵍⵣ	ⵜⵍⵜⵓⵙⵀⵜ ⵜⵣⵍⵜⵓⴰⵀ ⴰⵏⵓⴰⵡ ⴰⵙⵣⵀ

Tableau 1: les 10 plus proches voisins

Nous pouvons voir que le modèle construit donne des résultats impressionnants. il semble que les mots « ⵎⴰⵢⵓⵣⵙ», « ⴰⵎⵙⵓⵣⴻⴰⵍ» et « ⵏⵜⴰⵯⵓ» soient très proche dans l'espace des représentations.

Nous pouvons ensuite voir à nouveau les analogies formées par l'addition et la soustraction des représentations. Une analogie particulièrement intéressante que notre modèle apprend est celle de «pays => capital», comme indiqué dans la première colonne du tableau ci-dessous. Les notions de genre semblent aussi plus claires dans ce modèle, puisque les plus proches voisins de la représentation « homme + femme - roi = reine » en amazigh « ⴰⵓⵅⴰⵜ + ⵜⴰⵎⴻⴻⵙⴻⵜ - ⴰⵅⵉⵉⵣⵏ = ⵜⴰⵅⵉⵉⵣⵏⵜ».

D'après les résultats des tests, nous pouvons remarquer que leur modèle élaboré pouvait rapidement proposer des résultats assez prometteurs. Dans un premier temps, il a obtenu les mêmes matrices de représentation de vecteurs que l'on peut avoir avec les méthodes distributionnelles où il est très facile par exemple, d'associer ⵔⴰⵢⵙ avec ⵍⵣⴰⵢⵔ.

Il a donc amélioré la détection des synonymes de manière à ce qu'il puisse répondre à des questions du type, « quel est le mot qui est similaire à ⵜⴰⵔⴱⵉⵜ? » et ainsi créer des analogies. De là, il a pu augmenter le jeu de questions, de façon à obtenir aussi bien des relations sémantiques que syntaxiques. Le modèle a permis d'obtenir des résultats assez impressionnants, notamment certains permettant de réaliser des analogies entre une capitale et son pays.

4. Conclusion

Dans cet article, nous avons développé le modèle word2vec pour la langue amazigh en utilisant les données des pages web et les technologies Big Data. Nous avons fourni deux modèles; l'une basée sur Skip-gram et l'autre sur le modèle sac de mots continu. Pour montrer leur capacité à capturer la similarité entre les mots nous les avons évalués en utilisant plusieurs tests, sachant que l'utilisation des technologies Big Data nous a permis de gagner plus de 90% de temps d'exécution. Nous croyons que ces modèles élaborés peuvent être utilisés par d'autres chercheurs pour améliorer la performance de diverses tâches de PNL.

Références

Boukil, S., Adnani, F. E., Moutaouakkil, A. E. E., Cherrat L., Ezziyyani, M. (2017). Arabic Stemming Techniques as Feature Extraction Applied in Arabic Text Classification. in Advanced Information Technology, Services and Systems. pp. 349-361.

Samir, B., Mohamed, B., Fatiha, E. A., Loubna, C., Elmajid, E. M. (2018). Arabic text classification using Deep Learning technics. International Journal of Grid and Distributed Computing. Vol. 11, no 9.

Alayba, A. M., Palade, V., England, M. , Iqbal, R. (2018). Improving Sentiment Analysis in Arabic Using word Representation. arXiv:1803.00124.

Singh, S. K. , Paul, S., Kumar, D. (2014). Sentiment Analysis Approaches on Different Data set Domain: Survey.International Journal of Database Theory and Application. 7(5):39-50.

Youness, M., Mohammed, E., Jamaa, B. (2018).Semantic Indexing of a Corpus. International Journal of Grid and Distributed Computing. 11(7):63-80.

Al Qassem, L. M., wang, D., Al Mahmoud, Z., Barada, H., Al-Rubaie, A., Almoosa, N. I. (2017).Automatic Arabic Summarization: A survey of methodologies and systems. Procedia Computer Science. Vol. 117, pp. 10-18.

Mahdi, A. E., Alahmadi, A., Joorabchi, A. (2014).Combining Bag-of-words and Bag-of-Concepts Representations for Arabic Text Classification. in 25th IET Irish Signals & Systems Conference, Limerick, Ireland. pp. 343–348.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.(2013). Distributed Representations of w ords and Phrases and their Compositionality. arXiv:1310.4546.

Rong, X. (2014). word2vec Parameter Learning Explained. arXiv:1411.2738.

Karim, M. R. , Alla, S. Spark, S. (2017). For Big Data Analytics: Explore the concepts of functional programming, data streaming, and machine learning. Packt Publishing Ltd,.

Luu, H. (2018). Beginning Apache Spark 2: w ith Resilient Distributed Datasets. Spark SQL, Structured Streaming and Spark Machine Learning library. Apress.

Bhatnagar , K. Hart, B. (2015). Building Python Real-Time Applications with Storm. Packt Publishing.

Jain, A. (2017). Mastering Apache Storm: Real-time big data streaming using Kafka. Hbase and Redis. Packt Publishing Ltd.

TICAM'18, p. 127 - 138

Tifinaghee character recognition via structural features

Youssef OUADID, Brahim MINAOUI, Mohamed FAKIR
Department of Computer Science,Laboratory of Information Processing and Decision and Support,
Faculty of Science and Technology, Sultan Moulay Slimane University, Beni Mellal, Morocco
yo.ouadid@gmail.com,bra_min@yahoo.fr, m.fakir@usms.ma

Abstract

In this paper a system for the recognition of Tifinaghe characters is presented. It is divided into three main steps: preprocessing, feature extraction and classification. Image quality is enhanced through preprocessing techniques which are: thresholding, normalization and thinning. The image is given to a proposed key points extraction method, then transformed into a weighted undirected graph. The character is classified by matching the graph of the character and its counterpart graphs generated from IRCAM database using an efficient spectral graph matching algorithm. Experimental results are accomplished by the use of 3267 random characters to test the effectiveness. The system shows good results in term of accuracy and CPU time.

Keywords : OCR, Amazigh, Tifinaghe, Graph Matching, Key Points

1. Introduction

Optical Character Recognition system is an important tool for man-machine interaction. It is the process of converting a text image into readable and easy to modify text by computer or similar material. Many researches have been concentrated into creating an efficient Optical Character Recognition system, especially for Latina, Chinese (Shah *et al.*, 2013) and Arabic (Lorigo *et al.*, 2006) characters. It is used in several areas where the work is based on the text documents, mainly in office for indexing purposes, automatic archiving of documents and banks to facilitate the reading of the amounts of checks. In the other hand, the recognition of the Amazigh characters, called Tifinaghe, remains less explored.

Amazigh language (Ameur *et al.*, 2004) is the oldest writing of North Africa, it has more than 3000 years of existence. It is with the royal speech in 2001 that the legitimization of the Amazigh language is formalized. since then, the Royal Institute of the Amazigh Culture (IRCAM) have been marketing this language in order to safeguard, promote and strengthen the place of Amazigh culture in the educational, socio-cultural, national media space and in the management of local and regional affairs. It developed and adopted the Tifinaghe alphabet as the character of Amazigh language writing.

Since the introduction of this alphabet into the universal coding system in 2004, IRCAM's research centers have made their own efforts in numerous and in-depth studies on the promotion of this alphabet, the broadening of its radiation and its integration into information

systems. This resulted in the appearance of Amazigh documents in Tifinaghe printed or handwritten characters. As a result, the automatic processing of these documents has become a very active field of research.

Tifinaghe alphabet is composed of 33 graphemes corresponding to the 33 phonemes of the standard Amazigh. Table 1, presents these 33 characters as well as their Latin correspondents.

Amazigh script is written from left to right. Unlike Latin language, it does not have the notion of uppercase and lowercase, nor that of pseudo-word. However, it includes the same punctuation marks as the Latin one. It is a non-cursive writing which facilitates the process of segmentation. This justifies the fact that most of the research done on this writing is focused on character recognition since existing works on the segmentation of Latin script documents are viable for Amazigh.

ya	ⴰ	a	yaḥ	ⵃ	ḥ	yaṛ	ⵕ	ṛ
yab	ⴱ	b	yaɛ	ⵄ	ɛ	yaɣ	ⵖ	ɣ
yag	ⴳ	g	yax	ⵅ	x	yas	ⵙ	s
yag^{W}	ⴳⵯ	g^{W}	yaq	ⵇ	q	yaṣ	ⵚ	ṣ
yad	ⴷ	d	yi	ⵉ	i	yac	ⵛ	c
yaḍ	ⴹ	ḍ	yaj	ⵊ	j	yat	ⵜ	t
yey	ⴻ	e	yal	ⵍ	l	yaṭ	ⵟ	ṭ
yaf	ⴼ	f	yam	ⵎ	m	yaw	ⵡ	w
yak	ⴽ	k	yan	ⵏ	n	yay	ⵢ	y
yak^{W}	ⴽⵯ	k^{W}	yu	ⵓ	u	yaz	ⵣ	z
yah	ⵀ	h	yar	ⵔ	r	yaẓ	ⵥ	ẓ

Table 1: The official repertoire of the tifinaghe-ircam alphabet with their correspondents in latin characters

Here, a selective review of recently proposed Tifinaghe character recognition systems is presented. (Amrouch *et al.*, 2009) proposed an approach based on the extraction of directional information from the Hough transformation of each character in the form of a features vector. This information feeds a hidden Markov model (HMM). The results obtained are promising. However, the discrimination of these models is not very good, according to the authors,

when every letter is represented by a single reference image. To remedy this issue, (Amrouch *et al.*,2012) have replaced the Hough transform with a new technique to express a set of structural features from the contour of the character based on points that have maximum deviation. In the learning and classification phase, they combined dynamic programming with continuous HMM. This approach has the advantage of being independent of the number of recognition classes (in terms of memory and speed) since the model is built for all classes.

The results, which are quite encouraging, have shown that continuous HMM are more robust. However, the disadvantages of this approach are the detection of the points that seems restrictive for some fonts of the Amazigh writing. In an attempt to solve orientation and size change problems, (El Ayachi *et al.*, 2014) compared two robust descriptors. These are the invariant moments and the transform of w alsh. The authors presented two systems containing the same preprocessing and classification techniques. Using dynamic programming in the classification phase, the authors concluded that the invariant moments are greater than the w alsh transform in terms of execution time and discrimination. In order to improve the recognition rate, (El Ayachi *et al.*, 2011) have replaced the dynamic programming method with a single hidden layer neural network.

The system gave a better recognition rate compared to dynamic programming and neural networks with 2 or 3 hidden layers. Recently, in (El Ayachi *et al.*, 2014) the authors used their systems to display the braille code adapted to Tifinaghe characters. The braille system is a system adopted to help blind and partially sighted people integrate into different areas of life. (Es-Saady *et al.*, 2008) proposed a syntactical approach for the recognition of printed Tifinaghe characters. This is done by representing the character images using Freeman coding. The classification is performed using the finite automata. According to the authors, the tests carried out show the robustness of the system.

However, the problem with this approach is that it does not deal with circular characters. To remedy this, the authors used the horizontal and vertical symmetry of the spelling of the Tifinaghe alphabet. Indeed, the authors presented in (Es-Saady *et al.*, 2010) a system based on the position of the central lines of each character. The features are extracted based on the density of the pixels contained in a sliding window in the image of the character. According to the authors, this approach has proved its power of discrimination by testing it on a local database. In order to complement the limits of the previous systems facing the problems of rotation and size change, (Bencharef *et al.*, 2011) proposed an approach based on a geometric description using geodetic descriptors. These descriptors served as input to the hybrid classification process that combines neural networks and decision trees. The success of Bencharef's approach has motivated Oujaoura *et al.* to go in the same direction. In a first approach (Oujaoura *et al.*, 2013a), the authors performed a comparison between the Walsh transform, GIST and texture using Bayesian networks as classifiers. The tests of these descriptors on a local database showed the superiority of the GIST method in terms of recognition rate and computation time.

In a second approach (Oujaoura *et al.*, 2013b), the authors have proposed a system that combines Zernike moments, Legendre moments, Hu moments, walsh and GIST transforms in the feature extraction phase; As well as the neuron networks, SVM and NN in the classification phase. The results obtained are excellent in terms of recognition rate. However, the system is quite slow in terms of computation time. (Aharrane *et al.*, 2017) presented an end-to-end system devoted to automatic recognition of printed Amazigh script in complex document images containing different languages such as web images and natural scene images. The system shows great results in term of accuracy.

In this paper, we propose an algorithm for the extraction of feature/key points. The goal is to provide a minimum number of key points that will represent correctly the structure of the character which will provide a more accurate graph representation. This algorithm is included into an optical character recognition system. It contains three main steps: Pre-processing, Structural Feature Extraction and Classification.

The organization of the remainder of this paper is as follows: section 2 deals with the image enhancement techniques, which consists of binarization, normalization and thinning. In section 3, features extraction method is described. Section 4 deals with the classification method. In the last section, experimental results and discussion are given.

2. Pre-Processing

The pre-processing aims to reduce the number of data to keep only relevant information. It consists of: thresholding, normalization and thinning. The thresholding is done using the well-known Otsu method (Xu *et al.*, 2011).

2.1. Normalisation

Normalization removes the unwanted areas and put the character in the center of the image. To do it, the vertical and horizontal histogram of the image are calculated to find character borders. The horizontal histogram is scanned vertically and the vertical histogram is scanned horizontally from both direction to find the first white pixel in top, bottom, left and right. Figure 1(b) illustrate an example of “yah” character image normalization.

2.2. Thinning

Thinning aims to produce a pixel width version of the character while preserving its shape, connectivity, topology and end of route. we adopted Zhang-Suen algorithm (Zhang and Suen, 1984). It is a fast- parallel thinning algorithm that consists of two sub iterations. The algorithm explores all white pixel and remove the ones that do not belong to the skeleton.

By analyzing visually, the thinned database of Tifinaghe characters, the algorithm yields good result in term connectivity & shape preservation. Figure1(c) illustrate an example of “yah” character image thinning.

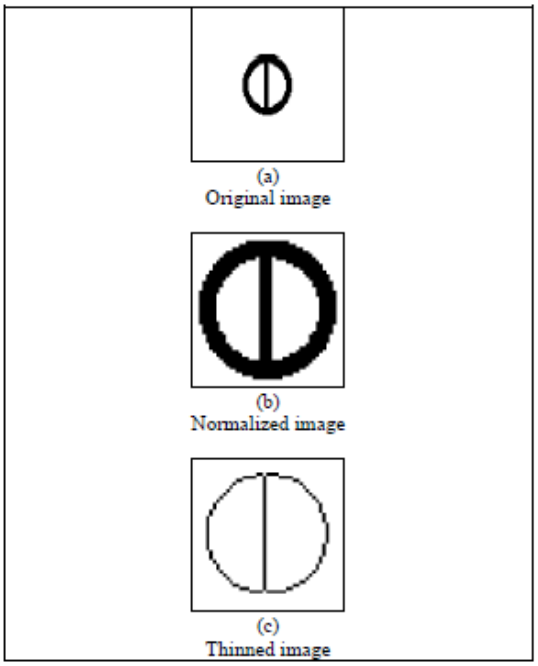

Figure 1: Example of "yah" character thinning and normalisation

2.3. Key points Extraction

The aim of this step, is to provide a synthetic description of the character. Using structural approach, which is graph representation. w e proposed an algorithm to divide the skeleton of the character into several straight-line segments. The algorithm divides the curved segment into minimum straight- line segments.

In order to correctly divide the character skeleton into several strokes and represent it with an undirected graph, authors started by extracting primary key points which are branch and end points. The next step is to find secondary key points. Figure 2 illustrate the proposed key points extraction process.

Using primary key points, the skeleton is segmented. Each segment is classified into a straight-line or curve segment classes. The classification is done using a threshold which compare the segment length with the Euclidian distance between its ends. The value of the threshold is calculated empirically. In this case, 0.2 provided the best recognition results. Hence, the segment is a curve is its length is more then 20% longer then the Euclidian distance of its ends.

$$D - d > 0.2*\mathrm{d} \qquad (1)$$

when a curve is found, the orthogonal distance between the elements of the segment and the Straight Line connecting its ends (SL) is calculated. The element E that have the maximum

orthogonal distance is chosen as a secondary key point. Every time a secondary point E is added into a segment, a new segment (SE) and (EL) are added to the list of segments and the old segment SL is removed from the list segments. The algorithm for key points extraction is as follow:

1. Extract primary key points.
2. Extract the list of segments based on those points.
3. For every segment, check if it is a curve.
4. For every curve, find the element that have the biggest orthogonal distance and design it as a key point.
5. Replace the segment with the new segments.
6. Repeat 3 to 5 till there is no curve in the segment list.

lgorithm output are a segments list where every segment is a set of pixels coordinates and the first and the last pixel are key points. An in-depth review of the presented in future work.

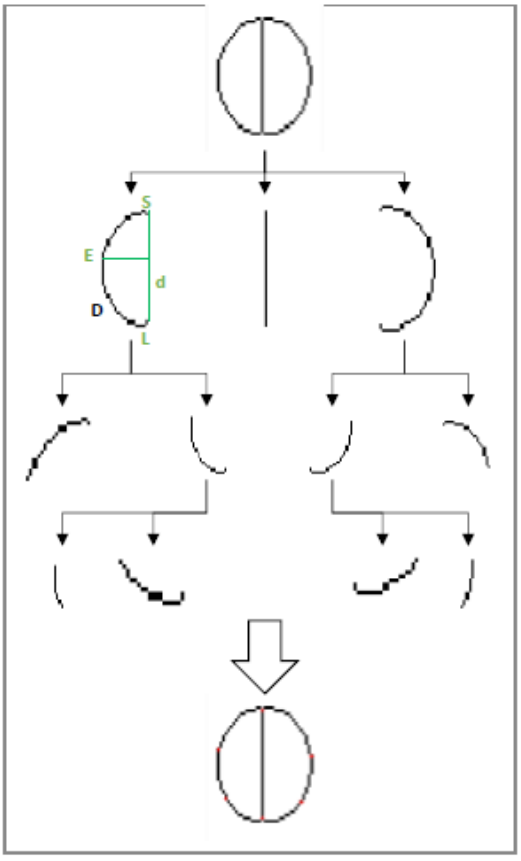

Figure 2: "yah" character key points extraction using authors proposed algorithm

2.4. Graph construction

Many works on pattern recognition by graph representation have been done (Vento, 2015). All of them are agreed to the fact that the conversion of an image into a graph is very important step. The graph should represent the characteristics of the letter to a large extent. A graph is a formal mathematical representation of a set of objects and their relationships. Each object is called a vertex. Relationships between objects are called edges. More formally, we define a graph G as an ordered pair of G = (V, E) where V is a set of vertices, E is a set of edges and each edge is a pair of vertices.

Adjacency matrix (AM) is a way to represent graphs. It is a binary square matrix M where the number of vertices |V| is its size. The entry in row *i* and column *j* is non-zero if and only if the edge (*i, j*) is in the graph which means:

AM (i, j) = 1, nodes i and j are linked by an edge

AM(i, j) = 0, nodes i and j are not linked by an edge

Some graphs can be directed which means AM(i, j) ≠ AM(j, i), or undirected which means AM(i, j) = AM(j, i).

After the character segmentation algorithm finishes, the features extracted will be represented by an undirected graph, where nodes represent the key points and edges represents segments connecting two. The algorithm of graph construction is proceeded as follow:

1. Let 1, 2 …*n* be labels of the key points, where *n* is the number of those points.
2. Let *AM* be an $n \times n$ adjacency matrix.
3. For every segment four information are extracted: first key point, second key point, length of the segment, orientation of the segment.
4. Based on those information, the adjacency matrix is constructed where:

$$\begin{aligned} &AM(i, j) = w, \text{ keypoint } i \text{ is connected to } j \\ &AM(i, j) = 0, \text{ else} \end{aligned} \tag{3}$$

where,

$$w = 2 \times \text{segment length} + \text{segment orientation} \tag{4}$$

Figure 3 and 4 presents an example of adjacency matrix and graph illustration of yaḍ character.

0	0	0	45	0	0
0	0	0	0	59	0
0	0	0	0	0	59
45	0	0	0	85	84
0	59	0	85	0	0
0	0	59	84	0	0

Figure 3: Weighted adjacency matrix of yaḍ character

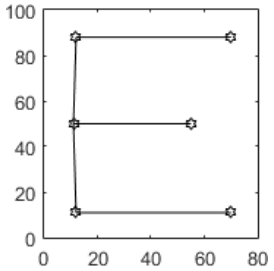

Figure 4: Illustration of yaḍ character graph using matlab gplot function

3. Classification

3.1. Spectral graph matching

The spectral method we used (Leordeanu and Hebert, 2005) aims to find coherent agreement between two sets of features P (P contain nP data features) and Q (Q contain nQ data features). It is generally used to find the main cluster in a graph. This is done by constructing the adjacency matrix M, where vertices represent the potential assignment (a = (i, i'), i ∈ P and i' ∈ Q) between two features and arcs design the level of resemblance between these assignments. This level is represented by a positive weight if two pairs of assignments are well agreed otherwise the link have zero weight. In our case, we used one-to-one correspondence mapping (instead of one-to-many). Which means that one feature of the first sets of features can be assigned with only one feature from the second set. This representation takes into account the geometry of the character and the degree of similarity between features. By observing of the adjacency matrix M, we can extract the spectral properties that can help us to determine how well a pair of assignments is connected to the main cluster. we keep rejecting correspondence with low association until we reach the one-to-one correspondence mapping. The proceeding to construct the matrix M is as follow: Given C a set of pairs where a = (i, i') ∈ C . It contains all possible assignments that respect the correspondence mapping one-to-one (one feature from P assigned at most to one feature from Q). To measure agreement between two features from P and Q, it is considered a list L of an associated score for each candidate assignment a ∈ C and an associated affinity for each pairs of assignments (a, b) where a, b∈ C. Based on this list we store these scores on the matrix M as follow:

- The diagonal entries of M contain the score of individual assignment a ∈ C.
- Other entries contain the score of pairs assignment (a, b) where a, b ∈ C.

The matrix M is a square matrix of size $n = k * np$ where k is the average number of candidate of each data features i ∈ P. Since we used interest points as features, all pairs of assignments are candidate assignment. Now the correspondence problem is reduced to find the cluster C that maximizes the inter cluster score.

$$S = \sum_{a,b \in C} M(a, b) = x^T M x \quad (5)$$

x is an indicator vector where:

$$x(a) = 1 \text{ if } a \in C \quad (6)$$
$$x(a) = 0 \text{ if } a \notin C$$

To find the best cluster C* we have to find the optimal binary vector x* that maximize the inter cluster score S.

$$x^* = \text{argmax}(x^T M x) \quad (7)$$

The value of this score depends on the number of assignments, the number of arcs adjacent to these assignments and the weight on arcs. The higher is the score S the bigger is the similarity between characters represented by M.

This method is applied by following the steps described in the algorithm below:

1. Build the symmetric non-negative n × n matrix M.
2. Initialize L by all possible candidate assignments and x a × n zeros vectors.
3. Find $a^* = \text{argmax}_{a \in L}(x^*(a))$, where x* is the principal eigenvector of M is.
4. If x*(a*) = 0 stop and return the solution x . Else, set
5. x(a*) = 1 and remove a* from L.
6. Remove all assignments in conflict with a* (one to one correspondence constraint).
7. If L is empty return the solution x. Else, return to step three & four.

4. Experimental results

Results are obtained from experiments on IRCAM Tifinaghe database (Ait Ouguengay and Taalabi, 2009). It is composed of 3300-character images with variation of size and fonts. In preprocessing phase, histogram method for normalization and Zhang-Suen algorithm for thinning are used. Figure 6 illustrates the computing time of preprocessing phase. Using IRCAM database, authors did not feel the need of a proper noise reduction algorithm since the noise level in images is low and Otsu method was more than enough to remove it. In feature extraction step, key point's extraction is very important to give a proper structural representation to the character. For every character a number of key points is desired as shown in table 2, but it is hard to get the exact number of key points considering changes of fonts and sizes in every character images. Figure 7 illustrates a comparison between the key points extracted per character via the proposed method and the desired key points. By analyzing figure 7&8, the algorithm we used give us very good results in term of key points/ CPU-time. In classification phase, 33 images are used as learning database and 3267 are used as test database. As a result, 22 images are misclassified, most of them are "Ya" and "Yar" character. when we looked into the database, we noted that there are some images of the two characters that are very similar even for the human eyes. w e compared our system with some of the best systems of Tifinaghe character recognition in the same device. The table 3 shows the recognition rate, error rate and CPU-Time obtained for each system.

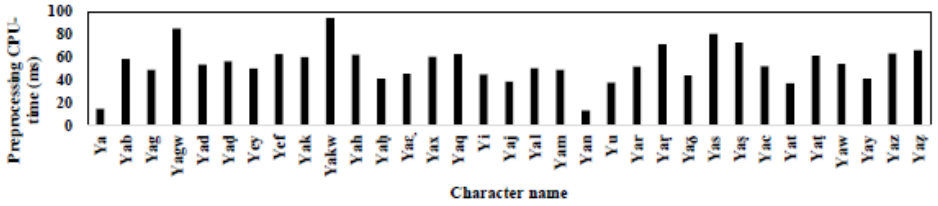

Figure 5: Preprocessing CPU-Time/Character id

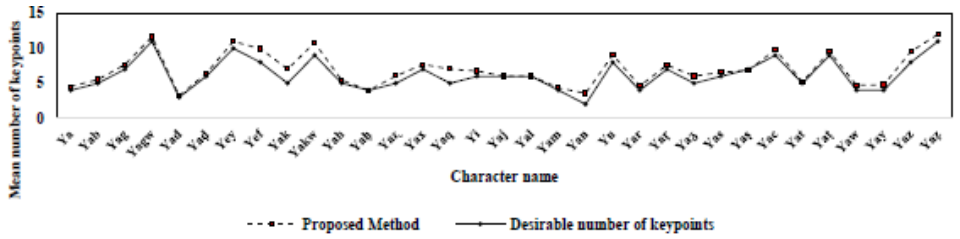

Figure 6 : Mean number of interest points per character

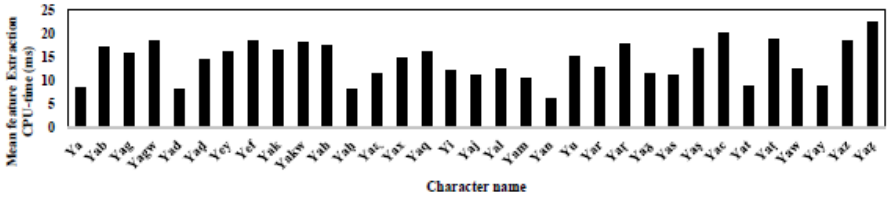

Figure 7: Mean CPU-Time of feature extraction per character

Ya	Yab	Yag	Yagw	Yad	Yad	Yey	Yef	Yak	Yakw	Yah	Yah	Yaç	Yax	Yaq	Yi	Yaj
4	5	7	11	3	6	10	8	5	9	5	4	5	7	5	6	6
Yal	Yam	Yan	Yu	Yar	Yar	Yaş	Yaş	Yac	Yat	Yat	Yaw	Yay	Yaz	Yaz		
6	4	2	8	4	7	5	6	7	9	5	9	4	4	8	11	

Table 2: Desirable key points for every character

Recognition system	Descriptor	Structural Feature Extraction	Gist Descriptor	Moments of Legendre
	Classifier	Spectral graph matching	Bayesian Network classifier	Multilayer Neural Networks
Recognition rate (%)		99	98	81
Error rate (%)		1	2	19
CPU Time (sec)		3200	5912.676	28 154.28

Table 3: Recognition system comparison

5. Conclusion

In this paper, we presented a fast and accurate system for the recognition of Tifinaghe character. It is based on matching graphs via an efficient method. It uses the spectral properties of the affinity matrix to calculate the degree of similarity between graphs. Every graph represents structural features of the character which are key points and their connectivity. The results of experiment show that most of character are recognized without compromising CPU- Time. In future works, we are considering to propose classification algorithm for better results especially in term of CPU-Time.

References

Shah, M., Jethava, G. B. (2013). A literature review on hand written character recognition.

Lorigo, L. M., Govindaraju, V. (2006). Offline Arabic handwriting recognition: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 28(5):712-724.

Ameur, M. , Bouhjar, A., Boukhris, F. (2004). Initiation à la langue amazighe. Institut Royal de la Culture Amazighe,.

Amrouch, M., Rachidi, A., El Yassa, M., Mammass, D. (2009). Printed amazigh character recognition by a hybrid approach based on Hidden Markov Models and the Hough transform. in Multimedia Computing and Systems. pp. 356-360.

Amrouch, M., Es-Saady, Y., Rachidi, A., El-Yassa, M., Mammass, D. (2012).A novel feature set for recognition of printed amazigh text using maximum deviation and hmm. Int J Comput Appl. Vol. 44.

El Ayachi, R., Oujaoura, M., Fakir, M., Minaoui, B. (2014). Code Braille et la reconnaissance d'un document écrit en Tifinaghe. In Proceedings of the International Conference on Information and Communication Technologies for the Amazigh,.

EL Ayachi, R., Fakir, M., Bouikhalene, B. (2011). Recognition of TIFINAGHE Characters Using A Multilayer Neural Network. Int. J. Image Process (IJIP). 5(2):109.

Es Saady, Y., Rachidi, A., Elyassa, M., Mammass, D. (2008). Une méthode syntaxique pour la reconnaissance de caractères Amazighes imprimés. CARI'08.

Es Saady, Y., Rachidi, A., El Yassa, M., Mammas, D. (2010). Printed Amazigh Character Recognition by a Syntactic Approach using Finite Automata. *Int. J. Graph. Vis. Image Process.* 10(2):1-8.

Bencharef, O., Fakir, M., Minaoui, B., Bouikhalene, B. (2011). Tifinagh Character Recognition Using Geodesic Distances, Decision Trees & Neural Networks. IJACSA Int. J. Adv. Comput. Sci. Appl. Spec. Issue Artif. Intell. pp. 1-5.

Oujaoura, M., Minaoui, B., Fakir, M. (2013). w alsh, Texture and GIST Descriptors with Bayesian Networks for Recognition of Tifinagh Characters. Int. J. Comput. Appl. Vol. 81, no. 12.

Oujaoura, M., El Ayachi, R., Minaoui, B., Fakir, M., Bouikhalene, B., Bencharef, O. (2013). Invariant descriptors and classifiers combination for recognition of isolated printed Tifinagh characters. in Third international symposium on Automatic Amazigh processing (SITACAM'13). Beni-Mellal, Morocco.

Aharrane, N., Dahmouni, A., Ensah, K. E. M., Satori, K. (2017). End-to-end system for printed Amazigh script recognition in document images. in Advanced Technologies for Signal and Image Processing (ATSIP). pp. 1- 6.

Xu, X., Xu, S., Jin, L., Song, E. (2011). Characteristic analysis of Otsu threshold and its applications. Pattern Recognition Letter. 32(7):956-961.

Zhang, T. Y. , Suen, C. Y. (1984). A fast parallel algorithm for thinning digital patterns. Commun. ACM. 27(3):236-239.

Vento, M. (2015). A long trip in the charming world of graphs for pattern recognition. Pattern Recognition. 48(2):291-301.

Leordeanu M., Hebert, M. (2005).A spectral technique for correspondence problems using pairwise constraints. in Computer Vision. Vol. 2, pp. 1482-1489.

Ait Ouguengay Y., Taalabi, M. (2009). Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe : Phase d'apprentissage. Systèmes Intelligents. -Théories Application.

Système de reconnaissance hors-ligne des caractères amazighes manuscrits basé sur les réseaux de neurones convolutifs profonds

Mohamed BENADDY[1], Othmane EL MESLOUHI[1]
Youssef ES-SAADY[2]
[1] LabSIE, Département de mathématiques, informatique et gestion,
Faculté Polydisciplinaire de Ouarzazate, Université Ibn Zohr
{m.benaddy,o.elmeslouhi}@uiz.ac.ma
[2] IRF-SIC, Faculté Polydisciplinaire de Taroudant & Faculté des sciences Agadir,
Université Ibn Zohr
y.essaady@uiz.ac.ma

Résumé

Avec l'intégration de la langue amazighe dans les technologies de l'information et de communication, la reconnaissance des caractères du tifinaghe est devenu un challenge pour les chercheurs. En raison de la diversité des styles des écritures, la reconnaissance des caractères tifinaghes manuscrits est une tâche complexe, d'où la nécessité de développement de systèmes performants pour la reconnaissance de ces caractères. Dans cet article, nous proposons un système hors ligne de reconnaissance des caractères tifinaghes manuscrits isolés. Ce système est basé sur les réseaux de neurones convolutifs profonds (Deep Convolutional Neural Networks) qui tire ses avantages du domaine de l'apprentissage profond (Deep Learning). L'extraction des caractéristiques se fait par un réseau de neurones de convolution à partir des images brutes. Cette technique n'exige pas les phases de prétraitement supplémentaires d'extraction de caractéristiques comme le font les techniques classiques. Ce système est testé sur l'intégralité de l'ensemble des caractères de la base de données AMCHD. Comparé avec d'autres techniques proposées dans la littérature, le système proposé donne de meilleurs résultats qui atteint un taux de reconnaissance de 99,10%.

Mots clés : *reconnaissance automatique des manuscrits ; réseaux de neurones, tifinaghe, l'apprentissage profond.*

1. Introduction

La langue amazighe est une langue afro-asiatique (wolff, 2018). Elle est parlée par des populations dispersées dans l'Afrique du Nord entre l'oasis de Siwa en Égypte, la Mauritanie, les îles Canaries et le nord du Sahara (Ameur *et al.*, 2004) La plus forte concentration des locuteurs amazighes se trouve au Maroc (wolff, 2018). La langue amazighe marocaine est divisée en trois variétés ; Tarifite dans le nord, Tamazight dans le centre et sud-est du Maroc et Tachelhit dans le sud-ouest et l'anti Atlas (Ameur et Souifi, 2004).

Avec le Discours Royal d'Ajdir (Khénifra) du 17 octobre 2001 (Ameur *et al.*, 2004) la langue amazighe commence à prendre sa stature puisqu'il a établit, par un dahir, la création et l'organisation de l'Institut Royal de la Culture Amazighe (IRCAM)[1], concrétisant l'annonce de sa fondation par le Roi Mohammed VI. En 2011, l'amazighe devient une langue officielle de l'état, en tant que patrimoine commun de tous les Marocains sans exception (Constitution, 2011). La langue amazighe a sa propre écriture depuis l'antiquité appelée tifinaghe (Es-Saady, 2012). Cette écriture est de nature alphabétique consonantique. Elle a déjà été utilisée depuis le sixième siècle avant l'ère du Christ par les populations d'Afrique du Nord, du Sahel et des îles Canaries (Es-Saady, 2012). Par cet alphabet sont écrites les anciennes inscriptions (voir figure 1) connues sous le nom de «Libyco-Berbère» trouvées en Afrique du Nord, dans le Sahara, la Méditerranée au sud du Niger, aux îles Canaries et à la frontière occidentale de l'Egypte (Ameur *et al.*, 2006).

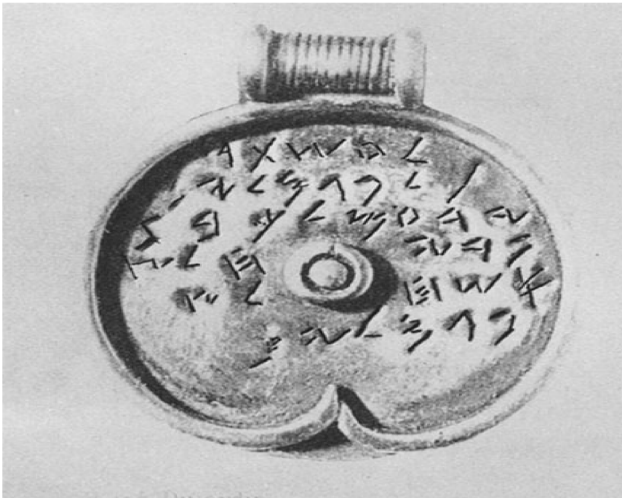

Figure 1: Médaillon de Carthage, source : (Casajus, 2013)

La graphie tifinaghe a été modifiée de ses origines à nos jours, du Libyque au Néo-tifinaghe, en passant par le tifinaghe saharien et le tifinaghe Touareg (Ameur *et al.*, 2004 ; Ameur *et al.*, 2006). La graphie Libyque est la plus ancienne utilisée dans la côte méditerranéenne de la Kabylie et au Maroc et probablement aux îles Canaries. Ce type est de forme occidentale, la forme orientale est utilisée en Constantine, dans les Aurès et en Tunisie (Es-Saady, 2012). Le tifinaghe saharien appelé aussi Libyco-Berber ou Touareg ancien, contient des signes supplémentaires, tels qu'une ligne verticale pour noter la voyelle finale /a/. Cette variété a été utilisée pour transcrire le Touareg ancien, mais ses inscriptions sont mal comprises. L'âge des inscriptions les plus récentes remonte probablement à environ 200 ans (Es-Saady, 2012). Le tifinaghe touareg, dans lequel sont incorporées quelques divergences dans la valeur attribuée aux signes qui correspondent aux variations du dialecte touareg, qui peuvent changer d'une région à l'autre (Es-Saady, 2012). Le Néo-tifinaghe, basé sur le tifinaghe

[1] http://www.ircam.ma/, visité en 2018.

Touareg, désigné pour l'écriture des dialectes amazighes du Maghreb (Maroc et Algérie) (Es-Saady, 2012 ; Ameur, 2006). Les anciens scripts tifinaghes sont gravés dans les pierres et les tombeaux de certains sites historiques du nord de l'Algérie, du Maroc, de la Tunisie, des régions touarègues du Sahara et des îles Canaries (Casajus, 2013). La figure 2 montre une image d'un ancien script tifinaghe trouvé sur le site de Dougga (nom actuel de l'ancien Thugga en Tunisie, connu sous le nom d'inscription du mausolée d'Atban Casajus, 2013). Plus d'informations sur l'ancien et le moderne tifinaghe peuvent être trouvées dans le livre (Ameur *et al.*, 2006) publié par l'IRCAM. Il fournit également une histoire de l'alphabet, ses origines, ses différentes variantes et leur déchiffrement.

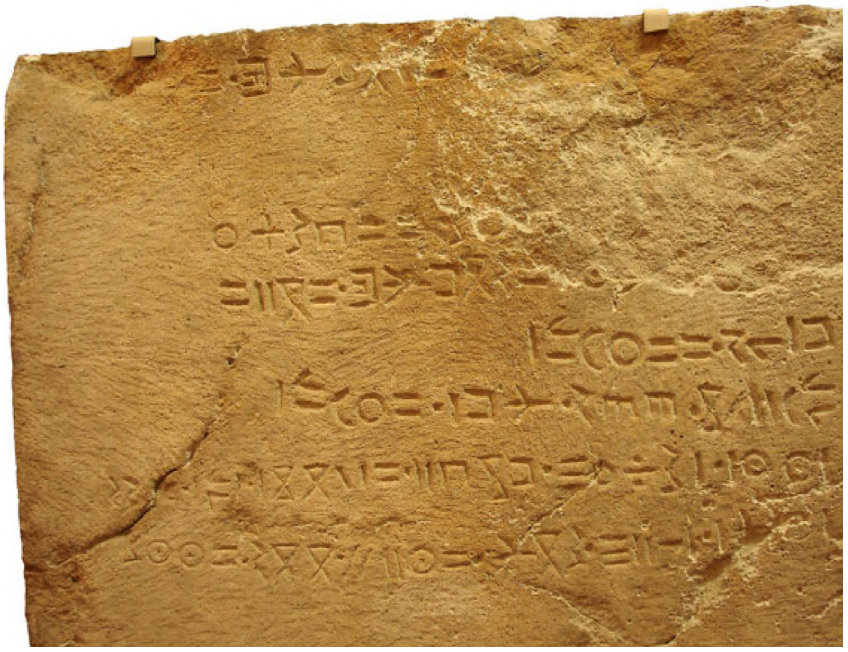

Figure 2: Inscription du Mausolé d'Atban (Libyque text), source :(Casajus, 2013)

La langue amazighe a été intégrée dans le système éducatif marocain depuis 2003. Elle est enseignée dans certaines classes des écoles primaires, dans la perspective d'une généralisation progressive (Boukhris *et al.*, 2008). Depuis le 10 février 2003, l'IRCAM a adopté le tifinaghe comme alphabet officiel de la langue amazighe et l'a appelé Tifinaghe-IRCAM (Ameur *et al.*, 2004). Comme initiative de standardisation, cet alphabet contient 33 caractères (voir figure 3). Le Tifinagh-IRCAM est officiellement reconnu par l'Organisation internationale de normalisation (ISO) (Zenkouar, 2004 ; Es-saady, 2014). La figure 3 montre les caractères tifinaghe-IRCAM reconnus et utilisés au Maroc avec leur prononciation en caractères latins.

Figure 3: Caractères tifinaghes adoptés par l'IRCAM avec leur prononciation en caractères Latin

Grâce aux progrès récents de la puissance informatique, plusieurs techniques de reconnaissance automatique de l'écriture manuscrite ont été développées, notamment pour les écritures latines et arabes (Byun et Lee, 2002 ; El Abed et Märgner, 2011 ; Koerich *et al.*, 2003 ; Marti et Bunke, 2002 ; Plötz et Fink, 2009 ; Tagougui *et al.*, 2013). Cependant, la grande variabilité inhérente à la nature de l'écriture manuscrite a rendu ce domaine de recherche très actif. Ainsi, ces dernières années, avec la croissance des moyens de communication, d'autres alphabets, tels que l'alphabet tifinaghe de la langue amazighe ont intégré les systèmes d'informations. Ce qui a entraîné l'apparition d'autres types de documents où l'écriture n'est pas encore traitée et donc plus attrayante à reconnaître. La reconnaissance automatique des caractères de ces documents nécessite des techniques de traitement plus spécifiques. Récemment, quelques efforts ont été rapportés dans la littérature pour la reconnaissance des caractères tifinaghes (Aharrane *et al.*, 2017 ; Aharrane *et al.*, 2015a ; Aharrane *et al.*, 2015b ; Amrouch *et al.*, 2012a ; Amrouch, 2012b ; Es-Saady *et al.*, 2011a ; Es-Saady *et al.*, 2011b ; Es-Saady *et al.*, 2011c ; Es-Saady *et al.*, 2014 ; Rachidi *et al.*, 2014). Dans cet article, nous présentons un réseau de neurones convolutifs profonds (Deep CNN) pour le système de reconnaissance des caractères tifinaghes manuscrits amazighes. Sur la base de notre connaissance de la littérature, cet article est le premier qui propose le modèle de réseau de neurones de convolution profond pour la reconnaissance des caractères amazighes manuscrits isolés. Le reste de cet article est organisé comme suit : La section 2 est consacrée à la revue de la littérature des avancées dans le domaine de reconnaissance des caractères amazighes manuscrits. La section 3 introduit les réseaux de neurones convolutifs profonds. La section 4 détaille le système proposé. La section 5 cite les résultats obtenus avec leur interprétation et la dernière section conclut l'article.

2. Revue de la littérature

Afin de développer et de tester des systèmes de reconnaissance des caractères amazighes manuscrits Es-saady *et al.* (Es-Saady *et al.*, 2011b) ont créé une base de données de caractères amazighes manuscrits appellée AMHCD (Amazigh Handwritten Character Database). La base AMCHD est la seule base de données disponible et d'une taille suffisante pour tester les systèmes de reconnaissance ainsi développés.

Dans leur article Rachidi *et al.* ont présenté un état de l'art et une comparaison des travaux de recherche scientifique accomplis et publiés dans le domaine de la reconnaissance automatique des caractères amazighes (Rachidi *et al.*, 2014). Dans (Aharrane *et al.*, 2015a), Aharrane *et al.* ont établi une étude comparative de différents algorithmes supervisés pour un système de reconnaissance des caractères amazighes manuscrits. Leur objectif est de comparer une liste de méthodes de classification populaires et de tester les performances de l'ensemble des caractéristiques extraites de caractères isolés en utilisant des méthodes statistiques avec ces différentes méthodes. Ils ont proposé dans (Aharrane *et al.* 2015b) un système de reconnaissance de caractères amazighes manuscrits basé sur une approche statistique avec un ensemble de 79 éléments. L'ensemble des caractéristiques élaboré comprend 37 entités de densités et 42 fonctions d'ombres basées sur un principe de zonage spécifique pour représenter les caractères amazighes. Dans la phase de reconnaissance, ils ont utilisé le perceptron multicouche (MLP) comme classificateur. Le taux de reconnaissance obtenu en utilisant 93,93% (24180) des caractères de la base de données AMHCD est de 96,47%, récemment, une amélioration de ce dernier système a été publiée dans (Aharrane *et al.*, 2017).

Amrouch *et al.* présentent un système automatique de reconnaissance de caractères amazighes basé sur les Modèles de Markov cachés (HMM) (Amrouch *et al.*, 2012a ; Amrouch, 2012b). Après des prétraitements sur l'image du caractère, une chaîne représentative du caractère est construite à partir de la transformation de Hough. La chaîne obtenue est traduite en séquence d'observations qui sera utilisée lors de la phase d'apprentissage par le modèle développé. La classification se fait ensuite par le calcul de la probabilité des séquences d'observations en utilisant l'algorithme de Viterbi (Augustin, 2001). Le modèle ayant le score maximum est retenu pour prédire la classe du caractère en entrée. Les auteurs ont évalué les performances de leur système sur 25740 caractères de la base AMHCD avec deux variantes. La première adopte la modélisation discrète des probabilités d'émission, quant à la deuxième utilise les MMCs continus. Ainsi, dans le cas continu, ils ont testé l'influence du nombre d'états et du nombre des gaussiennes sur le taux de reconnaissance. Le meilleur score obtenu est de 97,89% avec une topologie du nombre d'états égale à 14 et du nombre de gaussiennes égale à 1 ou à 2. Les résultats rapportés montrent l'efficacité de la modélisation proposée, notamment dans le cas continu, où les densités de probabilités sont modélisées par des gaussiennes.

Dans (Es-Saady *et al.*, 2011c), les auteurs ont proposé une nouvelle approche connexionniste basée sur la ligne centrale horizontale du caractère. La position de l'axe horizontal du caractère est utilisée pour dériver un ensemble de caractéristiques statistiques des pixels en utilisant la technique des fenêtres glissantes. Ces caractéristiques alimentent ensuite un réseau de neurones multicouches dans les phases d'apprentissage et de reconnaissance. Le réseau de neurones utilisé est composé d'une couche d'entrée de 90 neurones (correspondants aux caractéristiques), une couche cachée et une couche de sortie de 31 neurones (correspondants

au nombre de classes). Une amélioration de ce système a été publiée dans (Es-Saady *et al.*, 2011a ; Es-Saady, 2012), en intégrant d'autres caractéristiques basées sur l'axe centrale verticale du caractère afin d'exploiter les similarités des caractères tifinaghes selon cet axe. Pour l'expérimentation de cette approche, l'auteur a utilisé 24180 (780x31) caractères de la base AMHCD et la technique de validation croisée 10 fois. Les tests ont été effectués en fonction de l'intégration des caractéristiques dépendantes de la ligne centrale horizontale et verticale du caractère. Un taux de reconnaissance de 92,23% est enregistré lorsqu'il intègre les caractéristiques dépendantes de la position de l'axe horizontal et augmente à 94,62% lorsqu'il ajoute les caractéristiques dépendantes de l'axe verticale. Les causes des erreurs sont principalement dues à la ressemblance entre certains caractères amazighes. En fait, la déformation de l'écriture manuscrite influence la symétrie des caractères par rapport aux axes horizontaux et verticaux. Pour surmonter ces limitations, l'auteur a utilisé, dans (Es-Saady *et al.*, 2014), une ligne de base de référence variée sur le caractère au lieu de prendre la ligne centrale. L'étape d'extraction des primitives est précédée par l'estimation des lignes de base en se basant sur la méthode de projection et l'analyse des maxima et des minima du contour. Le taux de reconnaissance est augmenté à 94,96% lorsqu'il remplace les axes centraux du caractère par les lignes de base. Ce qui montre que les caractéristiques basées sur la position des lignes de base offrent une amélioration significative des performances de reconnaissance.

Toutes les approches citées précédemment utilisent un module de prétraitements et un autre d'extraction de caractéristiques avant la phase d'apprentissage et de classification. L'inconvénient majeur de ces approches est que la sélection des caractéristiques pertinentes qui constitue la phase la plus importante dans le processus de reconnaissance de caractères nécessite un temps de calcul important. En outre, les phases de prétraitement et d'extraction des caractéristiques ont également une influence directe sur les performances des classificateurs utilisés qui sont susceptibles de souffrir du problème de sur-apprentissage. Afin de remédier à ces inconvénients, nous proposons dans cet article une approche basée sur l'apprentissage profond (deep learning) qui permet d'éviter les problèmes liés aux prétraitements, extractions de caractéristiques et au sur-apprentissage.

3. Les réseaux de neurones convolutifs

Un réseau neuronal convolutif (CNN) est une concaténation d'une couche d'entrée, d'une couche de sortie et de plusieurs couches cachées. Comparé aux réseaux de neurones entièrement connectés, le nombre de paramètres de ce type de réseau est largement réduit par le partage des poids et des biais (Goodfellow *et al.*, 2016). Les CNNs sont particulièrement utilisés dans de nombreux domaines où les motifs présents dans les données sont invariants dans le temps et l'espace comme pour les images, les vidéos ou les sons.

La figure 4 illustre une architecture type d'un réseau de neurones convolutif, en utilisant une image RVB comme entrée. D'abord, une opération de convolution est appliquée à l'image d'entrée en utilisant k ($k = 1$ *sur la figure*) filtres (masques) pour chaque canal R, G et B. Ces filtres agissent comme des détecteurs de caractéristiques de l'image d'entrée originale. Ensuite, une fonction non linéaire ψ est appliquée au résultat de l'opération de convolution pour obtenir une carte dite d'activation (également appelée carte de caractéristiques). Chaque

couche est suivie d'une couche de sous-échantillonnage (pooling) pour réduire la taille de la carte d'activation et pour donner l'invariance aux petites translations locales. Enfin, une couche entièrement connectée ayant la fonction *softmax* comme fonction d'activation est utilisée dans la couche de sortie pour effectuer la classification.

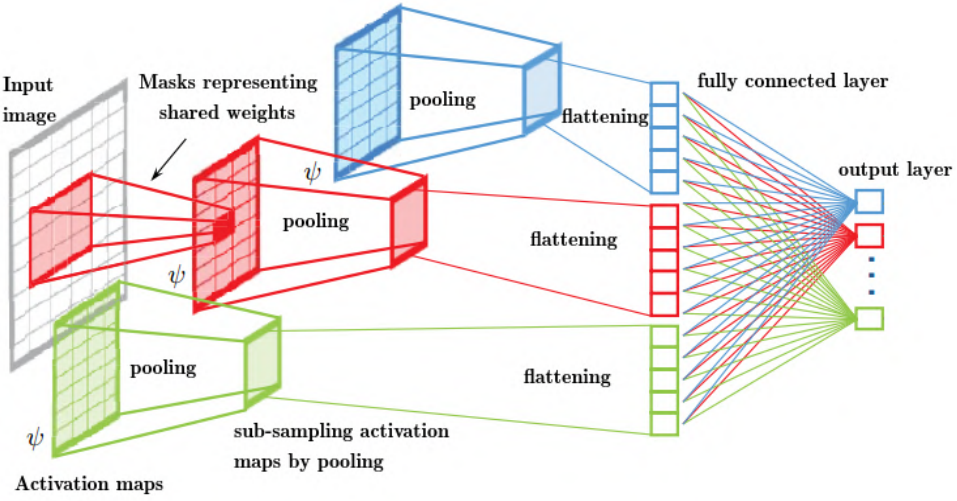

Figure 4 : Architecture type d'un RNC (CNN) avec une couche à convolution, une couche de maxpooling, une couche d'aplatissement (flattening) et une couche entièrement connectée avec un nombre fixe de neurones

4. Le système proposé

La figure 5 présente l'architecture du réseau adopté. Il est constitué de cinq couches adjacentes. Les trois premières couches (C1, C2 et C3) sont responsables de l'extraction des caractéristiques et les deux dernières (C4 et C5) effectuent la classification des caractéristiques. La première couche (C1) est la couche à convolution avec 32 cartes d'activation de 32 × 32 pixels chacune. Chaque neurone est associé à une convolution par un masque 3 × 3 et un biais. Dans cette couche, les différentes cartes d'activation correspondent aux différents masques et biais. Chaque carte a 9 poids plus le biais donnant 320 (32 × 10) paramètres à estimer durant l'apprentissage pour cette couche (L1). La fonction d'activation utilisée est l'unité linéaire rectifiée (Relu) dont la formule est définie par l'équation 1. Avant de présenter les images au réseau pour l'apprentissage, elles sont étiquetées et redimensionnées en 32 × 32 pixels.

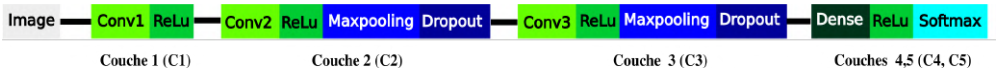

Figure 5 : L'architecture du système proposé

La couche 2 (C2) est composée d'une couche de convolution, d'une couche maxpooling et d'une couche de régularisation dropout. Elle utilise 32 cartes d'activation. La taille du pooling est de 2×2 dans les directions de x et de y. En utilisant le dropout, le pourcentage des neurones désactivés aléatoirement à chaque itération est de 50%. Au total, nous disposons de 9248 paramètres à apprendre pour cette couche.

La couche 3 (C3) est composée de trois couches : une couche à convolution, une couche maxpooling et une couche dropout avec 64 cartes d'activation. La taille de chaque carte d'activation est de 6×6 pixels. Au total, nous aurons 18496 paramètres à déterminer pendant la phase d'apprentissage pour la couche (L3). Cette combinaison des extracteurs de caractéristiques permettra l'exploration d'autres caractéristiques de haut niveau.

Les deux dernières couches (C4 et C5) effectuent la classification des caractéristiques. La couche 4 (C4) est composée de 64 neurones avec l'activation Relu. Chaque neurone de cette couche est entièrement relié à une seule carte d'activation de la couche (C3). La couche 5 (C5) contient 33 neurones indiquant la classe de l'image d'entrée avec une fonction d'activation softmax. Le système est optimisé à l'aide de l'optimiseur RMSProp avec un taux d'apprentissage adaptatif. Le nombre total de paramètres à estimer, durant l'apprentissage, de l'ensemble des couches de l'architecture proposée est de 177 729 paramètres.

$$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases} \tag{1}$$

5. Résultats

5.1. La base AMCHD

Pour évaluer l'approche proposée, nous utilisons la base de données Amazigh Handwritten Character Database (AMHCD). C'est une base de données de caractères amazighes manuscrits qui a été créée et développée au sien du Laboratoire IRF-SIC de l'Université Ibn Zohr d'Agadir (Es-Saady *et al.*, 2011b). La base contient 25740 images de caractères amazighes manuscrits isolés et étiquetés produits par 60 scripteurs de sexe, d'âge et de d'emplois différents. La figure 6 présente quelques exemples de caractères amazighes écrits à la main. Toutes les images de cette base de données sont redimensionnées à 32 × 32 pixels. Nous remarquons que contrairement aux méthodes existantes dans la littérature, notre système ne nécessite aucune étape de prétraitement des images de la base de données AMHCD.

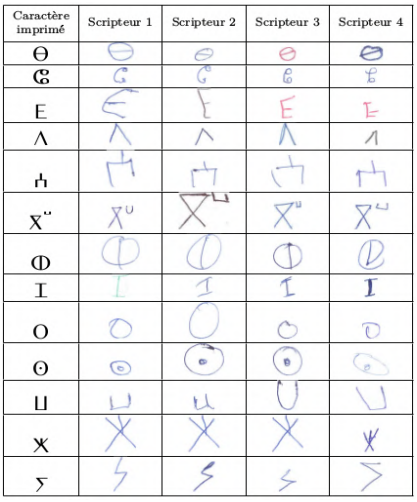

Figure 6 : Quelques caractères amazighes manuscrits de la base AMCHD

5.2. Résultats et discussions

Le système proposé est entraîné, d'une part, pour extraire les caractéristiques des caractères manuscrits et d'autre part pour obtenir les poids des différents neurones du réseau. L'ensemble de données est divisé en deux ensembles : l'apprentissage et la validation. A chaque itération, le réseau est entraîné par l'ensemble des images d'apprentissage. En utilisant les images de l'ensemble de validation, nous obtenons une estimation plus réaliste de la façon dont le modèle fonctionnerait avec des données invisibles et pour vérifier la présence d'un sur-ajustement. Les figures 7 et 8 montrent les courbes du taux de reconnaissance (accuracy) et de l'erreur qui est la différence entre le résultat obtenu par le système et celui attendu (loss) en fonction du nombre d'itérations. La figure 7 montre que notre réseau se comporte assez bien et atteint un taux de reconnaissance de 95%, après seulement 6 itérations. Comme nous pouvons le voir dans la figure 8, les données d'apprentissage et la validation continuent de chuter, avec de petits pics mineurs et aucun signe de sur-apprentissage.

Pour étudier les performances de notre système, nous utilisons des ensembles de nombre différent d'images durant les deux phases d'apprentissage et de teste. Le tableau 1 illustre les taux de reconnaissance obtenus en utilisant des tailles différentes des ensembles d'apprentissage et de teste. Comme nous pouvons le constater, le taux de reconnaissance atteint la meilleure valeur de 99,1% lorsque 80% des images de la base de données AMHCD sont utilisées pour l'apprentissage. Alors, le système est entraîné sur 80% des images et 20% qui reste pour le teste.

Taille de l'ensemble d'apprentissage	Taille de l'ensemble de validation	Taux de reconnaissance
50 %	50 %	94 %
60 %	40 %	98 %
70 %	30 %	99 %
80 %	20 %	99.1 %

Tableau 1 : Taille de l'ensemble d'apprentissage et la précision obtenue

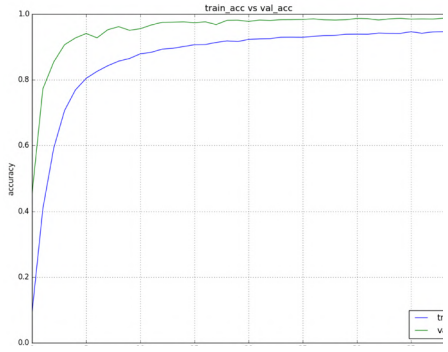

Figure 7 : Courbes de précisions durant la phase d'apprentissage (apprentissage et validation) obtenues par le CCN sur la base AMHCD

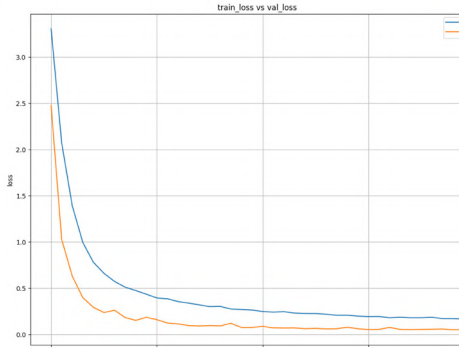

Figure 8 : La courbe de la perte (loss) durant la phase d'apprentissage (apprentissage et validation) obtenues par le CCN sur la base AMHCD

Pour plus de détails, la figure 9 donne un résumé des résultats des prédictions sous forme de la matrice de confusion. Comme on peut le constater, le nombre de caractères mal classés est très faible par rapport aux caractères bien classés (toutes les valeurs de la diagonale sont supérieures à 95%). Les erreurs commises par le système proposé ne sont pas surprenantes, elles peuvent s'expliquer par le fait que certains caractères ne sont pas bien écrits dans la base de données AMHCD.

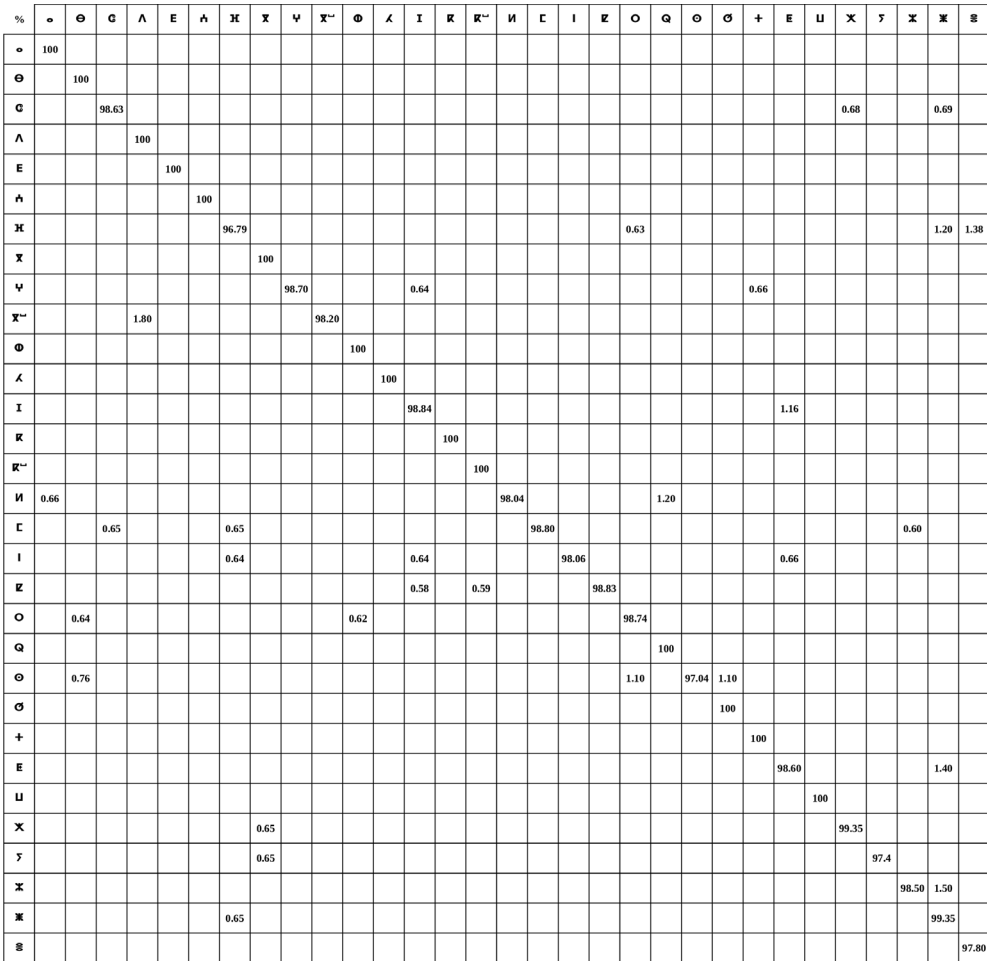

Figure 9 : Matrice de confusion obtenue sur l'ensemble des caractères de la base AMCHD

Pour prouver l'exactitude des résultats obtenus par notre système, ces résultats sont comparés avec ceux obtenus par diverses techniques existantes. Le tableau 2 donne les performances obtenues par la méthode proposée et par certains systèmes existants qui utilisent la base de données AMHCD. Comme nous pouvons le voir, l'approche proposée donne la meilleure performance sans aucune étape de prétraitement (comme dans (Aharrane, 2017 ; Aharrane, 2015b ; Djematene *et al.*, 1997 ; Es-Saady *et al.*, 2011a ; Es-Saady *et al.*, 2014), et donne la meilleure précision même si nous utilisons toutes les images de la base de données AMHCD (contrairement à toutes les méthodes citées dans le tableau 2).

Méthode	Taux de reconnaissance	Nombre d'images utilisées
Geometrical Method (Djematene *et al.*, 1997)	92.30 %	1700
Horizontal and vertical centerline of character (Es-Saady *et al.*, 2011a)	96.32 %	20150
Horizontal and vertical baseline of character (Es-Saady *et al.*, 2014)	94.96 %	24180
Continuous HMM and Directional features (Aharrane, 2015b)	97.89 %	20153
Combination of multiple classifiers with statistical feature (Aharrane, 2015a)	99.03 %	24180
Système proposé	99.10 %	25740

Tableau 2 : Comparaison des résultats obtenus par notre approche et d'autres méthodes dans la littérature

6. Conclusion

Dans cet article, nous avons présenté un système de reconnaissance des caractères amazighes manuscrits isolés, basé sur les réseaux de neurones à convolution profonde. Le système proposé opère directement sur les images brutes où tous les travaux publiés précédemment nécessitent de nombreuses étapes de prétraitement. Le système proposé a été évalué en utilisant l'ensemble des caractères de la base de données AMHCD et donne le meilleur taux de reconnaissance comparé aux autres systèmes existants dans la littérature. Dans les prochains travaux, nous avons l'intention d'étendre notre système de reconnaissance de phrases et de reconnaissance de l'écriture manuscrite multilingue.

Références

Aharrane, N., Dahmouni, A., El Moutaouakil, K., Satori, K. (2017). *A robust statistical set of features for amazigh handwritten characters*. Pattern Recognition and Image Analysis. 27(1):41-52.

Aharrane, N., El Moutaouakil, K., Satori, K. (2015). *A comparison of supervised classification methods for a statistical set of features: Application: Amazigh ocr.* In: Intelligent Systems and Computer Vision (ISCV). pp. 1-8.

Aharrane, N., Moutaouakil, K., Satori, K. (2015). *Recognition of handwritten amazigh characters based on zoning methods and mlp.* wSEAS transactions on Computers. Vol. 14, pp. 178-185.

Amrouch, M., Es-Saady, Y., Rachidi, A., El Yassa, M., Mammass, D. (2012). *Handwritten amazigh character recognition system based on continuous hmms and directional features.* International Journal of Modern Engineering Research. 2(2): 436-441.

Ameur, M., Bouhjar, A., Boukhris, F., Boukouss, A., Boumalk, A., Elmedlaoui, M., Iazzi, E., Souifi, H. (2004). *Initiation à la langue Amazighe*. Publications de l'Institut Royal de la Culture Amazighe, Manuels, No. 1.

Ameur, M., Bouhjar, A., Boukhris, F., Boukouss, A., Boumalk, A., Elmedlaoui, M., Iazzi, E., Souifi, H. (2006). *Graphie et orthographe de l'amazighe*, Publications de l'IRCAM, CAL, Rabat.

Amrouch, M. (2012). *Reconnaissance de caractères imprimés et manuscrits, textes et documents basée sur les modèles de Markov cachés*, Thèse de doctorat, Faculté des sciences d'Agadir.

Boukhris, F., Boumalk, A., El Moujahid, E. Souifi, H. (2008). La Nouvelle Grammaire de l'Amazighe, Centre de l'Aménagement Linguistique, Publications de l'IRCAM, Rabat.

Byun, H. and Lee, S. w. (2002). *Applications of support vector machines for pattern recognition: A survey.* In: (Ed.), Pattern recognition with support vector machines, Springer.

Casajus, D. (2013). *Sur l'origine de l'écriture libyque. Quelques propositions, Afriques (En ligne), Débats et lectures, mis en ligne le 04 juin 2013*, consulté le 28 février 2018. URL : http://journals.openedition.org/afriques/1203.

Dahir n° 1-11-91 du 27 Chaabane 1432 (29 juillet 2011) portant promulgation du texte de la Constitution », *Bulletin officiel du Royaume du Maroc*, ° 5964, 1902. ISSN 0851-1217.

Djematene, A.; Taconet, B., Zahour, A. (1997). *A geometrical method for printed and handwritten Berber character recognition*. Vol. 2, pp. 564-567.

El Abed, H., Märgner, V. (2011). *ICDAR 2009-Arabic handwriting recognition competition*, International Journal on Document Analysis and Recognition (IJDAR). Vol. 14, pp. 3-13.

Es-Saady, Y. (2012). *Contribution au développement d'approches de reconnaissance automatique de caractères imprimés et manuscrits, de textes et de documents amazighs*, Ibn Zohr University, Agadir.

Es-Saady, Y., Rachidi, A., El Yassa, M., Mammass, D. (2011). *Amazigh handwritten character recognition based on horizontal and vertical centerline of character.* International Journal of Advanced Science and Technology. 33(17):33-50.

Es-Saady, Y., Rachidi, A., Yassa, M., Mammass, D. (2011). AMHCD: *A database for amazigh handwritten character recognition research*, Int. J. Comput. Appl. Vol. 27, pp. 44-48.

Es-Saady, Y., Rachidi, A., El Yassa, M., Mammass, D. (2011). Reconnaissance automatique de lecriture amazighe à base de ligne centrale de l'écriture. 4ème Atelier international sur l'Amazighe et les TIC 2011.

Es-saady, Y., Amrouch, M., Rachidi, A., El Yassa, M. Mammass, D. (2014). *Handwritten Tifinagh Character Recognition Using Baselines Detection Features*, International Journal of Scientific & Engineering Research, Vol. 5, Issue 4.

Goodfellow, I.; Bengio, Y. Courville, A., (2016). *Deep Learning*. MIT Press. http://www.ircam.ma/ (2018) , accessed on Fabruary 20, 2018.

Koerich, A. L.; Sabourin, R. Suen, C. Y. (2003). *Large vocabulary off-line handwriting recognition: A survey*, Pattern Analysis & Applications. Vol. 6, pp. 97-121.

Marti, U-V., Horst, H. (2002). *The IAM-database: an English sentence database for offline handwriting recognition*, International Journal on Document Analysis and Recognition. Vol. 5, pp. 39-46.

Plötz, T., and Fink, G. A. (2009). *Markov models for offline handwriting recognition: a survey*, International Journal on Document Analysis and Recognition (IJDAR). Vol. 12, pp. 269.

Rachidi, A., Eddahibi, M., Es-saady, Y., Amrouch, M. (2014). *Amazigh Characters Automatic Recognition: Overview and Prospects*, International Journal of Scientific & Engineering Research. 5(11):797-803.

Tagougui, N., Kherallah, M.,Alimi, A. M. (2013). *Online Arabic handwriting recognition:* a survey, International Journal on Document Analysis and Recognition (IJDAR). Vol. 16, pp. 209-226.

w olff, H. E. (2018). Berber languages, accessed on Fabruary 20, 2018 https://www.britannica.com/topic/amazigh-languages.

Zenkouar, L. (2004). *L'écriture Amazighe Tifinaghe et Unicode*, Etudes et documents berbères. Paris (France). N°22, pp. 175-192.

TICAM'18, p. 153 - 166

Printed Tifinagh script recognition from Web and natural scene images in multilingual environment

Nabile AHARRANE[1], Abdellatif DAHMOUNI[1], Karim EL MOUTAOUAKIL[2], Khalid SATORI[3]

[1] University Sidi Mohammed Ben AbedAllah
[2] National school of applied sciences Al-Hoceima
[3] University Sidi Mohamed Ben Abed Allah
{aharranenabil, abdellatifdahmouni}@gmail.com
yassirkarimimane@gmail.com khalidsatori@gmail.com

Abstract

In this work, we present an end-to-end system devoted to automatic recognition of printed Amazigh script in complex document images containing different languages such as web images and natural scene images. To this end, text extraction from images is performed. The extracted text serves as input for a trained convolutional neural network (CNN) to identify its language. Finally, we proceed to the recognition of the Amazigh text script using a developed optical character recognition (OCR) system. The CNN reaches 99.12% of accuracy, while the OCR system gets 99,93%. The obtained results seem to be very satisfactory.

Keywords: Amazigh OCR; language identification; deep learning; convolutional neural network; machine learning.

1. Introduction

The contextual information in images of web pages and the natural scene images taken by the cameras integrated in different mobile devices are more and more interesting in the field of computer vision. The embedded text in this kind of images is considered as very important for image understanding which makes its processing a very important task for researchers due to its various applications such as images indexing, documents sorting, translation and augmented reality.

On the other hand, the images may contain multilingual text making the identification of the text language a primordial task to determine the correct recognition system to use.

In Morocco, after the formalization of the Amazigh language in 2003, the Amazigh script is increasingly used in various fields such as literature, Web, advertising, education and many others and consequently in web images and natural scene images (Figure 1). This requires an automatic processing. Unfortunately, all works dedicated to the Amazigh script recognition supposed that the document under processing is an Amazigh document where the text is extracted beforehand and does not handle multilingual documents.

In this work, we were interested in the recognition of the printed Amazigh script present in web images and natural scene images. Thus, we propose an end-to-end system composed of three main steps. First, text detection in images and its extraction is performed, and then the extracted text serves as input for a trained convolutional neural network CNN to identify the text script language. Finally, we proceed to the recognition of the Amazigh text script using a developed OCR system based on statistical methods to extract the character features that distinguish each character from the others.

Figure 1: The use of the Amazigh script in different areas.
(a) A Moroccan classroom of Amazigh teaching;
(b) An Amazigh TV channel;
(c) A traffic sign;
(d) A touristic advice.

The rest of this paper is organized as follows: In Section 2, we give a presentation of the Amazigh language and its characteristics. The proposed approach is detailed in Section 3. Section 4 reports the obtained experimental results. w e conclude the paper with Section 5.

2. The Amazigh language

The Amazigh language has existed since the earliest antiquity. It has an original writing system, called Tifinagh, used and preserved to this day. In recent decades, all Amazigh groups have reclaimed this ancestral writing. Currently, the Amazigh language is spoken by about 30 million speakers in North Africa (from the Oasis of Siwa in Egypt to Morocco passing through Libya, Tunisia, Algeria, Niger, Mali, Burkina Faso and Mauritania). In Morocco, where nearly 50% of people are Amazigh, the Amazigh language is divided into three regional varieties with Tarifit in the North, Tamazight in Central Morocco and South-East, and Tachelhit in South-w est and the High Atlas (Ameur *et al.*, 2014).

Figure 2: The Tifinagh characters adopted by the IRCAM.

The official introduction of the Amazigh language's teaching in the Moroccan educational system in 2003 involved the selection of a standard common language to teach. This task was accomplished by "The Royal Institute of the Amazigh Culture" (IRCAM) created in July 2001. Actually, the Tifinagh-IRCAM alphabet is based on 33 characters (Figure 2). In the Amazigh OCR field, the characters ⵅ⵿ and ⵇ⵿ do not have Unicode codes, so we obtain them by a combination of characters 'ⵅ, ⵇ' with the sign of labialization '⵿' that have Unicode codes.

3. Proposed approach

The proposed system is a top to down pipeline that extracts text blocks from input web images or natural scene images, identifies the text language, looks for the Amazigh one and proceed to its recognition. Figure 3 shows the necessary steps to construct the system.

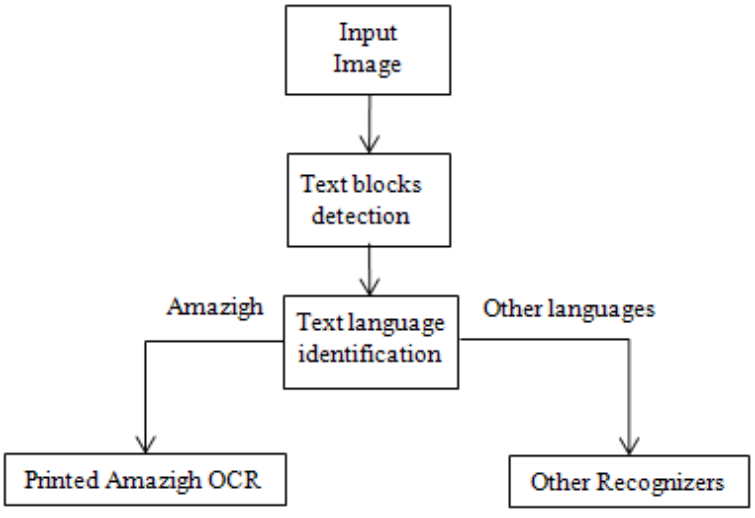

Figure 3: The Adopted System

This section presents in details the tree main steps used to create this system.

3.1. Text blocks Detection

The text present in the image is considered as an important aspect to understand the image, so its detection and extraction has become a key problem. This operation is primordial to prepare the text for the OCR task. Many works have been carried out in this topic, an excellent survey of the widely used algorithms in this field is presented in (Zhang *et al.*, 2013). Most of these algorithms are based on edge, texture and connected-component methods.

In this work, we use the edge-based text regions detection algorithm proposed by (Liu and Samarabandu, 2007).The authors proposed an algorithm that takes in input a web image or natural scene image and produces the boundary coordinates of blocks enclosing text by performing three stages: candidate text region detection, text region localization and character extraction. This algorithm showed its robustness in the face of variations of font, size, style, orientation, color/intensity and other complexities.

3.2. Text script identification

web images or natural scene images contain text of different languages which requires the identification of the text script to well orient the text block to the adequate OCR system. This operation has also other applications like indexing and sorting such images. However, few works were devoted to this context; some of the recent works are presented in (Brodić *et al.*, 2016; Ghosh *et al.*, 2010; Lu, 2015). No works have been carried out for identifying the Amazigh script in images against others languages. In Morocco, the three used language are Amazigh, Arabic and French. So, we aim to create a system able to classify the text blocks obtained from the previous step to the corresponding language. The system must identify also the numbers. To this end, due to its performance, we train a convolutional neural network (CNN).

CNN is one of the most known architectures of neural networks. It's a key algorithm of deep learning which is subject of intense research. Some of recent applications of CNN are given in (Krizhevsky *et al.*, 2012; Simonyan and Zisserman, 2014; Zeiler and Fergus, 2013). The CNN is a sequence of layers where each layer is composed from neurons of learnable weights. Every layer transforms one volume of activations to another through a differentiable function. The three layers most used to build a CNN are convolutional, pooling and fully-connected layers.

- Convolutional layer: The convolutional layer is the basic building block of a CNN. Its main role is to extract features from the input image. Using convolution, small squares of the input image participate in learning the features maps, so the spatial information is preserved.
- Pooling layer: This layer allows the compression of the image information by reducing the size of the resulted convoluted images using a simple sub-sampling.
- Fully-connected layer: This layer looks like layers used in the multilayer perceptron. It aims to connect every neuron from the max-pooled layer to every one of the output neurons.

3.2.1. Database and preprocessing

To train the used CNN, we developed a local database consisting of 7200 word images for each class; a total of 28800 images. 66.66 % of the database (19200) was subject of training set and 33.33 % for test (9600). Some of the samples used for training are given in Figure 4.

To create the database, we followed several steps:

- First, we collected 1800 different words for each of the four classes. For the Amazigh language, words were extracted from some books such as a French-Amazigh-Arabic dictionary and a Media dictionary published by the IRCAM institute. Arabic and French words were chosen from two files of the most commonly used words collected from movies subtitles. These files are under different formats and free downloadable from web site: http://www.101languages.net. whereas, the numbers were integers generated randomly between 1 and 20000.
- The collected words were rendered by the same rendering procedure used for the APTI database (Slimane *et al.*, 2012). Each word was rendered four times using different font, size and style to guarantee a large variety in text styles.

Figure 4: Samples of the word images in the training set

- The last step is a preprocessing step to make all the database images in a size of 50x50 that much the size of the CNN input image. This operation is automatic which takes into consideration the ratio between the width and the height of the word image and tries to duplicate the word, if needed as many times as necessary, so that the final image is a square image (Figure 4). Then the image is resized to 50x50 using a spline based algorithm (Muñoz *et al.*, 2001).

3.2.2. Learning

The created database serves to train a CNN composed of two convolutional layers of 20 and 10 feature maps respectively, both layers with 5x5 patch size and 2x2 max pooling. The last layer is a fully connected layer with four neurons that holds the scores for each class. The used CNN architecture is illustrated in Figure 5.

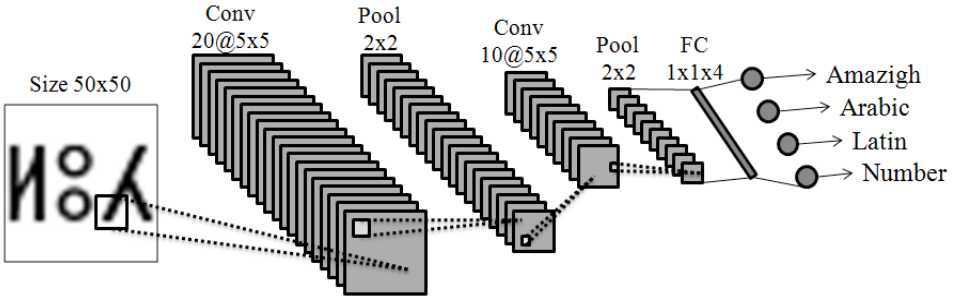

Figure 5: Architecture of the used Convolutional Neural netowrk

The CNN was trained 19200 epochs with a learning rate of 10^{-4} and the stochastic gradient descent as learning method.

3.3. Optical character recognition OCR

The main objective of an OCR system is to convert into readable and editable format different document images. In recent years, the OCR has received much attention in academic areas. It remains one of the most popular research subjects due to its diverse applications in the field of computer vision and image understanding. Therefore, much work has been achieved for many languages, an excellent and recent survey can be found in (Mathew *et al.*, 2016). Recently, researchers have begun to give attention to the Amazigh language. Few works were dedicated to Printed Amazigh OCR; the best performing ones are given in (Amrouch *et al.*, 2012; Ouadid *et al.*, 2016; Oujaoura *et al.*, 2013; Oujaoura *et al.*, 2014).

In this work, we were interested in the printed Amazigh Script. After the identification of Amazigh Text blocks in the image, we proceed the optical character recognition step (OCR). All the works proposed for the printed Amazigh OCR don't handle text presenting different fonts, sizes and styles which is the case for the texts in w eb images or natural scene images. Also, train another CNN can be useful due to its performance, but it is not recommended for images containing blocs of long text seen its computational time. To overcome these drawbacks and benefit of the fastness of statistical features, we use our recognition system proposed for Amazigh handwritten characters (Aharrane *et al.*, 2017) in the unique difference of changing the training and testing sets by printed Amazigh characters with different fonts, sizes and styles.

3.3.1. Preprocessing

This step consists of preparing the image for the next steps. To this end, multiple operations are performed:

- Thresholding: Otsu's method still one of the most used thresholding methods in character recognition systems (Otsu, 1979).
- Skew correction: consists of searching a rotation transformation such a way that makes the words baseline parallel to the horizontal (Aradhya *et al.*, 2006).

- Segmentation: The characters segmentation is one of the most important steps in an OCR system. The objective is to decompose the image into a sequence of sub-images in such a way that each sub-image must contain a single character.
- Normalization: The segmentation process produces isolated characters in different sizes. To solve this problem, a normalization step is needed. This latter consists on resizing all characters to a common size. In this work, due to its zooming quality, we used a spline-based algorithm (Muñoz *et al.*, 2001) to resize all characters to a size of 30x30.

3.3.2. Features extraction

As mentioned previously, in this step we use the same features extraction method used in (Aharrane *et al.*, 2017). This latter rests on decomposing the isolated character image into several overlapped zones according to different directions, then the density of black pixels and the total length of the histogram projection are calculated in each zone. The resulting set of features contains 79 components that will serve as input for different classifiers.

3.3.3. Recognition

After the extraction of features, we proceeded to the recognition phase, where we used three of the most used methods in the context of supervised classification that are the MultiLayer Perceptron (MLP), the Support Vector Machine (SVM) and the Random Forest (RF). These methods have shown their performance in different applications of pattern recognition and also when used in OCR systems.

Furthermore, the combination of multiple classifiers using majority voting in recognition phase is now considered as a very powerful tool to reach best performances. It had a great interest in characters' recognition research area. Its simple principle consists in giving to each instance the class that receives the largest number of votes by the classifiers. The use of the output probabilities of each classifier allows to this technique to have different variants to predict the output class of the combined classifier by using a combiner function such as the average, product, maximum and median rules (Rahman *et al.*, 2002).

Database:

In order to evaluate the performance of the proposed set of features, a local database was used as a source of training and test. This database consists of 6144 images of the isolated Amazigh printed characters generated with an automated procedure using different Amazigh fonts, sizes and styles for each character.

All the database images were created using a java program consisting in rendering the 32 Amazigh characters using 8 different Amazigh fonts (TABLE 1). These fonts are proposed by the IRCAM institute and are the most used fonts in the Amazigh writing. For each font, we used different sizes: 14, 16, 18, 20, 24 and 36 points. For each font and size combination, we used different styles: Plain, Bold, Italic and Bold-Italic combination. Thereby, with this rendering procedure, we obtained 6144 images.

Trainning:

Several experimentations have been conducted for all the three algorithms with different configurations under a compatible HP ProBook, Intel (R) Core (TM) i5-2520M CPU 2.50 GHz, and 4 GB of RAM through Java language. For the MLP, we carried out a series of runs by varying the number of neurons in the hidden layer. This number must be selected to be high enough to model the problem but not very high to avoid the overfitting. The experimentations for the SVM classifier were carried out with different kernel functions (Linear, Gaussian RBF and polynomial). As for as the RF, we execute the algorithm by varying in each run the number of trees constructing the forests.

Amazigh Word	**Font Name**
ⵜⵉⴼⵉⵏⴰⵖ	Tifinaghe-Ircam Unicode
ⵜⵉⴼⵉⵏⴰⵖ	Tifinaghe-Agoug_unicode
ⵜⵉⴼⵉⵏⴰⵗ	Tamzwart Standard UNICODE
ⵜⵉⴼⵉⵏⴰⵗ	Tamalout Standard UNICODE
ⵜⵉⴼⵉⵏⴰⵖ	Tifinaghe-taromit_unicode
ⵜⵉⴼⵉⵏⴰⵗ	Teddus Standard UNICODE
ⵜⵉⴼⵉⵏⴰⵗ	Tassafout Standard UNICODE
ⵜⵉⴼⵉⵏⴰⵗ	Tazdayt Standard UNICODE

Table 1: The used Amazigh fonts

In order to improve the system performance, we combine all of these classifiers using the majority vote system with different combiner functions such as Average, Product, Max, and majority.

All the classifiers were trained with the whole database using cross-validation 10-fold and were evaluated using the commonly accepted performance evaluation measures such as accuracy, recall and F-measure (Sokolova *et al.*, 2006).

Results: The obtained results for the different classifiers are reported in the following tables. Each table contains the obtained results for each classifier when running with different configurations.

Number of hidden nodes	Accuracy (%)	Recall (%)	F-Measure (%)
80	98,69	98,7	98,7
85	**98,86**	**98,9**	**98,9**
90	98,75	98,7	98,7
95	98,68	98,7	98,7

Table 2: Results for different Number of hidden nodes

Number of trees	Accuracy (%)	Recall (%)	F-Measure (%)
250	98,13	98,1	98,1
500	98,25	98,3	98,3
750	**98,27**	**98,3**	**98,3**
1000	98,25	98,3	98,3

Table 3: Results for different Number of trees

Kernel Function	Accuracy (%)	Recall (%)	F-Measure (%)
Linear	98,44	98,4	98,4
Gaussian RBF	98,68	98,7	98,7
Polynomial	**98,87**	**98,9**	**98,9**

Table 4: Results for SVM with different kernel functions

Voting Method	Accuracy (%)	Recall (%)	F-Measure (%)
Majority Vote	**99,93**	**99,9**	**99,9**
Max of probabilities	99,77	99,8	99,8
Average of probabilities	99.91	99,9	99,9

Table 5: Results of different classifiers combination methods

It should be noted that for the other languages we can use one of the performing OCR systems proposed in the literature.

4. Results and discussions

This section is devoted to present the APwID statistics and its corresponding information about storage and usefulness.

This section will summarize the obtained results for the different steps constructing the end-to-end proposed system.

For text detection, the used algorithm shows its robustness when dealing with Amazigh text presented in images. Figure 6 shows some results for images containing Amazigh text.

Figure 6: Results of detection algorithm [4] for images containing Amazigh script.

For the language identification step, the trained CNN was tested on a test set of 9600 word images; 2400 images for each class. The CNN reaches an accuracy of 99.12%, which we guess very satisfactory but presents an inconvenient because input word images extracted from images need a preprocessing step, as described above, to match CNN input size. Table 6 shows the obtained confusion matrix for the trained CNN.

	Amazigh	**Arabic**	**French**	**Number**
Amazigh	**2375**	6	16	3
Arabic	3	**2392**	3	2
French	30	5	**2349**	16
Number	0	0	0	**2400**

Table 6: Confusion matrix for the CNN

The classification errors seen in table VI are due to words containing common symbols between the different languages, especially for short words or words containing a single character. Table 7 shows some of the misclassified words in the test set.

Word	Class	Predicted class
ⵍⵛ	Amazigh	Latin
ⵄⵜ	Amazigh	Arab
ⵖⵏ	Amazigh	Number
أما	Arab	Amazigh
لا	Arab	Latin
زال	Arab	Number
FIN	Latin	Amazigh
Y	Latin	Arab
B	Latin	Number

Table 7: Examples of misclassified words in the test set

Finally, for the results of the proposed OCR system, Table 8 shows a global comparison of different used classifiers, combined with the used set of features, retaining only the best performance for each one.

Classifiers	Accuracy (%)	Recall (%)	F-Measure (%)
MLP	98,86	98,9	98,9
RF	98,27	98,3	98,3
SVM	98,87	98,9	98,9
Classifiers Combination	99.93	99,9	99,9

Table 8: Results for different classifiers

To showcase our method, we compared our obtained results with the other best performing approaches in term of Accuracy (Table 9).

Approaches	Accuracy (%)	Dataset size
Our Approach	**99,93**	**6144**
Ouadid *et al.*	99,02	3300
Oujara *et al.*	98,79	165
Oujara *et al.*	98.18	165
Amrouch *et al.*	97.50	240

Table 9: Comparison of the OCR system with other approaches

w e point that the obtained accuracy is very satisfactory and is the best one till now, in our knowledge, compared to other approaches in printed Amazigh OCR.

5. Conclusion

w e presented in this paper an end-to-end system able to detect text in web images and natural scene images. The detected text serves as input for a trained CNN classifier to identify its language and then proceed to its recognition by the adequate OCR system. The CNN classifier gave a good accuracy of 99.12%. To recognize the Amazigh language we proposed a printed OCR system that reaches 99.93%, which we guess very satisfactory. The language identification step presents an inconvenient seen that it needs a preprocessing operation to adapt the size of the extracted text image to the size of input images of the CNN.

Références

Aharrane, N., Dahmouni, A., El Moutaouakil, A., Satori, K. (2017). "A Robust Statistical Set of Features for Amazigh Handwritten Characters." *Pattern Recognition and Image Analysis.* 27(1): 41-52.

Ameur, M. *et al.* (2014). *Initiation À La Langue Amazighe*. IRCAM.

Amrouch, M. *et al.* 2012. A Novel Feature Set for Recognition of Printed Amazigh Text Using Maximum Deviation and HMM. *International Journal of Computer Applications.* 44(12): 23-30.

Manjunath, A. V. N. Kumar, G. H., Shivakumara, P. (2006). "Skew Detection Technique for Binary Document Images Based on Hough Transform." *Int. Journal of Information Technology.* 3(1): 194–200. https://pdfs.semanticscholar.org/3708/48c90fde1bdc4b0fa287acefece0da2eba34.pdf.

Brodić, D., Amelio, A., Milivojević, Z.N. (2016). Language Discrimination by Texture Analysis of the Image Corresponding to the Text. *Neural Computing and Applications.* 29(6): 151-72.

Ghosh, D., Dube, T., Shivaprasad, A. P. (2010). Script Recognition – A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. XX(12):1-21.

Krizhevsky, A., Sutskever, I., Hinton G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems. Vol. 1, NIPS'12, USA: Curran Associates Inc. pp. 1097-1105. http://dl.acm.org/citation.cfm?id=2999134.2999257.

Liu, X., Samarabandu, J. (2007). An Edge-Based Text Region Extraction Algorithm for Indoor Mobile Robot Navigation. *In IEEE International Conference Mechatronics and Automation*. pp. 2043-50.

Lu, S. *et al.* (2015). Scene Text Extraction Based on Edges and Support Vector Regression. *International Journal on Document Analysis and Recognition (IJDAR)*. 18(2):125-135. http://link.springer.com/10.1007/s10032-015-0237-z.

Mathew, M., Singh, A. K., Jawahar, C. V. (2016). Multilingual OCR for Indic Scripts. Proceedings - *12th IAPR International Workshop on Document Analysis Systems*, DAS. pp. 186-91.

Muñoz, Arrate, Thierry Blu, and Michael Unser. (2001). Least-Squares Image Resizing Using Finite Differences. *IEEE Transactions on Image Processing*. 10(9):1365-1378.

Otsu, Nobuyuki. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*. 9(1):62-66. http://ieeexplore.ieee.org/document/4310076/.

Ouadid, Y., Minaoui, B. Fakir, M. (2016). Spectral Graph Matching for Printed Tifinagh Character. *Proceedings-Computer Graphics, Imaging and Visualization: New Techniques and Trends*. pp. 105-11.

Oujaoura, M. *et al.* (2013). Application of Data Mining Tools for Recognition of Tifinagh Characters. *International Journal of Advanced Computer Science and Applications*. pp. 2-5.

Oujaoura, M., Minaoui, B., Fakir, F., El Ayachi, R., Benchare, O. (2014). Recognition of Isolated Printed Tifinagh Characters. *International Journal of Computer Applications*. 85(January):1-13.

Rahman, A.F.R., Alam, H., Fairhurst, M.C. (2002). Multiple Classifier Combination for Character Recognition: Revisiting the Majority Voting System and Its Variations. *Document Analysis Systems*. pp. 167-178. http://link.springer.com/chapter/10.1007/3-540-45869-7_21.

Simonyan, K., Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *In 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. pp. 1-14.

Slimane, F., Kanoun, S., Alimi, A. M. (2012). Database and Evaluation Protocols for Arabic Printed Text Recognition.

Sokolova, M., Japkowicz, N., Szpakowicz, S. (2006). Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. *In Advances in Artificial Intelligence*, eds. Abdul Sattar and Byeong-ho Kang. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 1015-1021.

Zeiler, M. D., Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *CoRR* abs/1311.2. http://arxiv.org/abs/1311.2901.

Zhang, H., Zhao, K., Song, Y. Z., Guo, J. (2013). Text Extraction from Natural Scene Image: A Survey. *Neurocomputing*. Vol. 122, pp. 310-323. http://dx.doi.org/10.1016/j.neucom.2013.05.037.

Amazigh alphabet speech recognition via IVR service

Mohamed HAMIDI[1,2], Hassan SATORI[1,2], Ouissam ZEALOUK[1,2], Khalid SATORI[1]

[1]LIIAN, Department of Mathematics and Computer Science, FSDM, USMBA, Fez, Morocco

Mohamed.Hamidi1@usmba.ac.ma

[2]Laboratory Artificial Intelligence Complex Systems and Modeling, Department of Mathematics and Computer Science FPN, UMP Nador, Morocco

hsatori@yahoo.com

Abstract

This article presents an evaluation of the Amazigh alphabets performance via an Interactive Voice Response (IVR) system based on the G711 audio codec. In this work, we study the effect of voice coding and decoding on the speech recognition rate based on Hidden Markov models (HMMs) and Gaussian Mixture Models (GMMs). In our approach to construct the VoIP network, we use the Asterisk server, Amazigh alphabets speech recognition system. The best system performance is found for the G711 codec, 3 HMM, and 16 GMMs.

Keywords: Automatic speech recognition, Amazigh language, IVR service.

1. Introduction

Interactive Voice Response (IVR) is a promising technology which automates interactions between callers and phone systems by pressing digits on the telephone or speaking words or short phrases. It enables the user to retrieve/enter information from a database using their voice in real time. The IVR consisted of several technologies working together such as computer telephony, speech recognition, to schedule, receive, enter, and record automated phone calls. This will allow an efficient exchange of information with reducing costs (Asterisk, 2015).

(Shah *et al.*, 2012) have studied a VoIP network using the asterisk server. The system was configured with different security parameters like VPN server, Firewall IP table rules, Intrusion Detection and Intrusion Prevention System. Their proposed system architecture was implemented using VMware. The researchers in (Anwar *et al.*, 2006) have examined the various VoIP attacks and it is prevent policies according to NIST report. The authors (Rafique *et al.*, 2009) have considered the DoS attack by categorizing the network into SIP dependent performance matrix and SIP independent matrix in order to evaluate the performance of VoIP.

(Basu *et al.*, 2012) have described the real-time challenges to design telephonic Automatic Speech Recognition System. In their study, they have used the Asterisk server to design a system which poses some queries and the spoken responses of users are stored and transcribed

manually for ASR system training. In this work, the speech data are collected from w est Bengal.

(Aust *et al.*, 1995) have created an automatic system that permits users to ask for train traffic information using the telephone. This system connects 1200 German cities. The caller can retrieve information talking fluently with the system which behaves like a human operator. The important components of their system are speech recognition, speech understanding, dialogue control and speech output which is executed sequentially.

(Bhat *et al.*, 2013) created the Speech Enabled Railway Enquiry System (SERES) which is a system that permits users to get the railway information considering the Indian scenario, as a case study to define issues that need to be fixed in order to enable a usable speech-based IVR solution.

(Satori *et al.*, 2014) have created a system based on HMM (Hidden Markov Models) using the CMU Sphinx tools. The aim of this work is the creation of automatic Amazigh speech recognition system that includes digits and alphabets of Amazigh language. The system performance achieved was 92.89 %. In (hamidi *et al.*, 2016) we present our first Experiment to integrate the ten first digits of Amazigh language in an Interactive Voice Response (IVR) server where the users use speech (ten first Amazigh digits) to interact with the system.

In this article, we compare the VoIP Amazigh ASR system performance by varying the values of their respective parameters as codecs, HMMs, and GMMs in order to determine the influence of codecs on the systems recognition rates.

The rest of this paper is organized as follows: Section 2 presents an overview of the VoIP system and protocols. Section 3 gives an overview of automatic speech recognition system. In section 4 Amazigh language. Section 5 Telephony Amazigh speech recognition will be discussed. Finally, section 6 is Experimental results. We finish with some conclusion.

2. VoIP System and protocols

VoIP (Voice over Internet Protocol) is a technology during last decade. It provides audio and video streaming facility on successful implementation in the network.

2.1. Asterisk

Telephony Server Asterisk is an open source and a development environment for various telecommunication applications programmed in C language. It provides establishment procedures enabling to manipulate communication sessions in progress. Asterisk supports the standard protocols: SIP, H.323 and MGCP and transformations between these protocols. It can use the IAX2 protocol to communicate with other Asterisk servers (Spencer *et al.*, 2003; Madsen *et al.*, 2011).

2.2. Session Initiation Protocol

The Session Initiation Protocol (SIP) is a signaling protocol which is responsible for creating media sessions between two or more participants. SIP was defined by Internet Engineering

Task Force (IETF) and is simpler than H.323 and adapted more specifically for session establishment and termination in VOIP (Handley, 1999). In our word SIP was used to create the user account and to assure Internetwork Communication.

2.3. Real-Time Transport Protocol

The Real-Time Transport Protocol (RTP) is an Internet protocol allows transmitting real-time data such as audio and video. RTP is a protocol which facilitates the transport of data over a network in real-time applications. It is intended to be used for applications such as audio and video conferencing, real-time systems control, and unicast or multicast services (Schulzrinnz, 1996).

2.4. Codec

The codecs are basically different mathematical tools used for encoding or compressing the analog voice signal into digital bit streams and back. The various codecs are based on an algorithm of compression, data rate and the sampling rate (Karapantazis, 2009).

3. automatic speech recognition system

3.1. *ASR system*

Speech recognition is the process of decoding the speech signal captured by the microphone and converting it into words (Huang *et al.*, 2001). The recognized words can be used as commands, data entry or application control. Recently, this technology has reached a higher level of performance. The applications of speech recognition found in several domains like healthcare, military, commercial/industrial applications, Telephony, Personal Computers and many others devices. Figure 1 shows the speech recognition system structure.

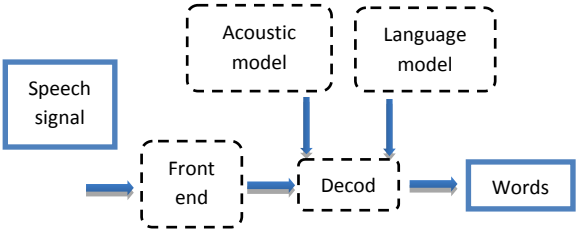

Figure 1: ASR system.

3.2. Hidden Markov Model

The Hidden markov model (HMM) (Beal *et al.*, 2002) is a popular method in machine learning and statistics for modeling sequences like speech. This model is a finite ensemble of states, where each set is associated with a probability distribution. Transitions among the states are governed by a set of probabilities called transition probabilities. Markov models are

excellent ways of abstracting simple concepts into a relatively easily computable form. Used in data compression to sound recognition. The Markov model makes the speech recognition systems more intelligent. In Table 1 we present the steps of a sequence generated by HMM. Below we present a example of 4 states HMM (See Figure 2).

T ‹— 1
Choose the initial state $q_1 = s_i$ with the probability π_i
while (t <= T)
{
Choose the observation $o_t = v_k$ with the probability bi(k)
Go to the next state q_{t+1} = sj with the probability a_{ij}
t ‹— t+1
}

Table 1: The hmm generator

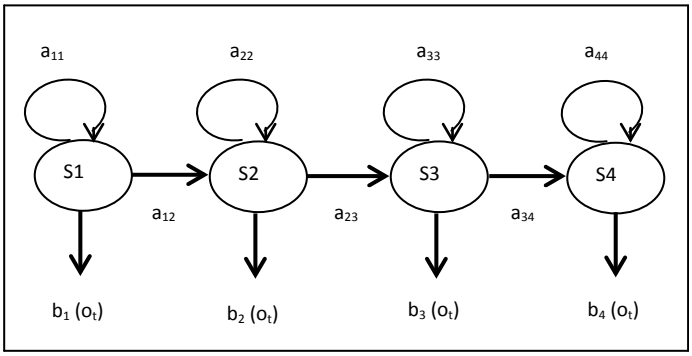

Figure 2: Example of HMM- 4 states

4. Amazigh language

Before the implementation of a Speech Recognition Voice Response Server System for any language, it is necessary to have a preliminary study of this language. In our case, we choose Amazigh which is a less-resourced Moroccan language. In the best of our knowledge, this is the first IVR using this language.

The Amazigh language is widely spoken in a vast geographical area of North Africa. It's spoken by 28 % of the Moroccan population. Table 1 represents the 33 alphabets of Amazigh language which are consecrated by IRCAM with their syllables and their transcription in English and Arabic.

Tifinagh	English transcription	Arabic transcription	Syllabes	No. Of syllabe
ⴰ	YA	يا	CV	1
ⴱ	YAB	ياب	CVC	1
ⴳ	YAG	ياڤ	CVC	1
ⴳⵯ	YAGG	ياڤ	CVC	1
ⴷ	YAD	ياد	CVC	1
ⴹ	YAḌ	ياض	CVCC	1
ⴻ	YEY	ياي	CVC	1
ⴼ	YAF	يااف	CVC	1
ⴽ	YAK	ياك	CVC	1
ⴽⵯ	YAḲ	ياك	CVCC	1
ⵀ	YAH	ياه	CVC	1
ⵃ	YAḤ	ياح	CVC	1
ⵄ	YAAA	ياع	CVC	1
ⵅ	YAX	ياخ	CVC	1
ⵇ	YAQ	ياق	CVC	1
ⵉ	YI	يي	CV	1
ⵊ	YAJ	ياج	CVC	1
ⵍ	YAL	يال	CVC	1
ⵎ	YAM	يام	CVC	1
ⵏ	YAN	يان	CVC	1
ⵓ	YU	يو	CV	1
ⵔ	YAR	يار	CVC	1
ⵕ	YAṚ	يارً	CVC	1
ⵖ	YAGH	ياغ	CVC	1
ⵙ	YAS	ياس	CVC	1
ⵚ	YAṢ	ياص	CVC	1
ⵛ	YASH	ياش	CVC	1
ⵜ	YAT	يات	CVC	1
ⵟ	YAṬ	ياط	CVC	1
ⵡ	YAw	ياو	CVC	1
ⵢ	YAY	ياي	CVC	1
ⵣ	YAZ	ياز	CVC	1
ⵥ	YAẔ	ياز	CVCC	1

Table 2: The 33 alphabets of Amazigh language

5. Telephony Amazigh Speech Recognition

In this section, we describe our experience to create and develop a Telephony Amazigh voice recognition system based-alphabets using Asterisk server and CMU Sphinx tools.

The system is created using Oracle virtual box tool on the host machine with 2 GB of RAM and an Intel Core i3 CPU of 1.2 GHz speed. The operating system used in our experiment was Ubuntu 14.04 LTS.

5.1. Speech Preparation

In this section, we describe our speech recognition database. The corpus consists of 33 spoken Amazigh letters collected from 30 Amazigh Moroccan native speakers aged between 16 and 50 years-old.

. The audio data are recorded in wave format by using the recording tool w avesurfer. Each alphabet is pronounced 10 times. In our work, the database is partitioned to training 70% and testing 30% in order to ensure the speaker independent aspect. More technical details about our system are shown in Table 3.

Parameters	Values
Sampling rate	16 kHz
Number of bits	16 bits
Audio format	w AV
Number of speakers –Training	21
Number of speakers –Test	9
Token number	9900
HMMs	3-5
GMMs	8-16-32-64

Table 3: the System parameters

5.2. ***Dictionary***

The pronunciation dictionary or lexicon includes all Amazigh alphabets we want to train followed by their pronunciation. The dictionary is the intermediary between the Acoustic Model and Language Model. Table 4 presents the used dictionary which includes the Amazigh alphabets.

5.3. Training Phase

In order to determine their optimal values for maximum performance, different acoustic models are prepared by varying HMMs (3- 5) and GMMs (8-16-32-64). The wave recorded audio data is used in the training phase where the database is partitioned to 70% training

and 30% testing in order to ensure the speaker independent aspect. More details about our training systems are shown in Table 5.

5.4. Telephony Recognizing Phase

Our idea is to acquire and process audio signals in real time from the transferred audio stream where the system was tested by coding audio data based G711 VoIP audio codec. The prepared system includes two major threads: a coding-decoding audio stream and the alphabets speech recognition process.

YA	Y A
YAB	Y A B
YAG	Y A G
YAGG	Y A GG
YAD	Y A D
YADD	Y A DD
YEY	Y E Y
YAF	Y A F
YAK	Y A K
YAKK	Y A KK
YAH	Y A H
YAHH	Y A HH
YAAA	Y A AA
YAX	Y A X
YAQ	Y A Q
YI	Y I
YAJ	Y A J
YAL	Y A L
YAM	Y A M
YAN	Y A N
YU	Y U
YAR	Y A R
YARR	Y A RR
YAGH	Y A GH
YAS	Y A S
YASS	Y A SS
YAC	Y A C
YAT	Y A T
YATT	Y A TT
YAw	Y A w
YAY	Y A Y
YAZ	Y A Z
YAZZ	Y A ZZ

Table 4: The first dictionary part used in the training

HMM states	GMMs	System description
3	8	Alphabets system using 3 HMMs and 8 GMMs
	16	Alphabets system using 3 HMMs and 16 GMMs
	32	Alphabets system using 3 HMMs and 32 GMMs
	64	Alphabets system using 3 HMMs and 64 GMMs
5	8	Alphabets system using 3 HMMs and 8 GMMs
	16	Alphabets system using 3 HMMs and 16 GMMs
	32	Alphabets system using 3 HMMs and 32 GMMs
	64	Alphabets system using 3 HMMs and 64 GMMs

Table 5: DIFFERENT TRAINING SYSTEMS

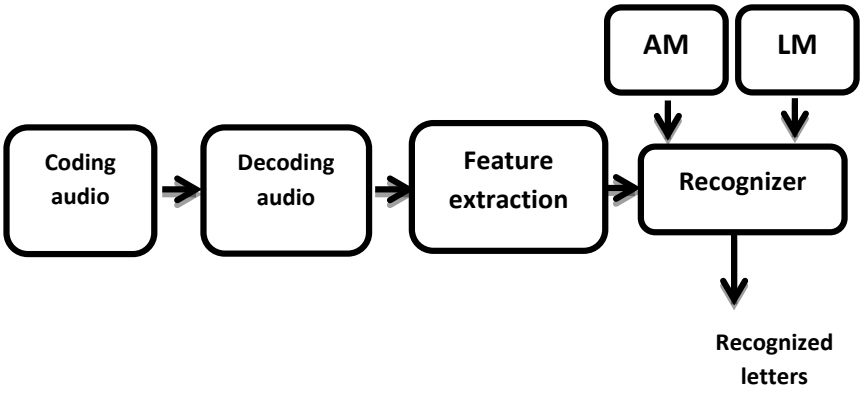

Figure 3: The scheme of the telephony Alphabets Amazigh speech recognition.

In the first step, the audio is transferred via IVR service. In the next step, the audio signal is split into frames and the MFCCs are calculated for each of them. These MFCCs are expressed with GMMs parameters and compared with the stored database. The recognition rate for each alphabet was observed and recorded for each experiment. Figure 3 presents the main telephony ASR system components.

6. Experimental results

In order to evaluate the performance of speech recognition system through the IVR server and to study the effect of VoIP audio codec on the recognition rate, we performed two majors experiments. In our experiments, the system was trained by using the recorded audio data and tested by using transferred audio data via IVR. Our system was trained using different HMMs and GMMs. The numbers of HMMs were 3 and 5 and Gaussian mixtures per model were 8, 16, 32 and 64. The obtained results are shown in Table 6.

In considering testing results it was found that, the most frequently recognized Amazigh alphabet is “YAQ” by using 16 GMMs. For 5 HMM-states the performances of the system are 80.67, 83.47, 78.25 and 77.47%. Our results show that the best recognition rate is 85.76 % achieved by 3 HMMs and 16 GMM.

In addition, it is noted that the system performance in all system experiments was better for 3 HMMs and 16 GMMs but lower for 5 HMMs and 64 GMMs.

By comparing our results with those of [8], we found that the voice coding has an effect on the recognition rates since all recognition rates achieved by using uncoded speech are higher than achieved by using coded-decoded speech. Table 7 presents overall system recognition rate for different parameters.

Alphabets	**3 HMMs**				**5 HMMs**			
	8	**16**	**32**	**64**	**8**	**16**	**32**	**64**
ⵄ	80,00	84,44	80,00	82,22	76,67	81,11	76,67	78,89
ⵖ	78,89	81,11	75,56	74,44	76,67	78,89	73,33	72,22
ⵀ	77,78	85,56	76,67	77,78	77,78	85,56	76,67	77,78
ⵃ	80,00	81,11	75,56	73,33	78,89	80,00	74,44	72,22
ⵍ	80,00	86,67	76,67	75,56	75,56	82,22	72,22	71,11
ⵎ	78,89	83,33	78,89	81,11	75,56	80,00	75,56	77,78
ⵏ	83,33	83,33	80,00	78,89	83,33	83,33	80,00	78,89
ⵓ	87,78	90,00	85,56	85,56	84,44	86,67	82,22	82,22
ⵔ	80,00	83,33	80,00	80,00	77,78	81,11	77,78	77,78
ⵕ	86,67	86,67	83,33	82,22	84,44	84,44	81,11	80,00
ⵙ	86,67	87,78	83,33	82,22	85,56	86,67	82,22	81,11
ⵛ	78,89	83,33	78,89	78,89	78,89	83,33	78,89	78,89
ⵜ	78,89	82,22	78,89	77,78	78,89	82,22	78,89	77,78
ⵡ	90,00	90,00	82,22	76,67	84,44	84,44	76,67	71,11

ⵣ	87,78	91,11	82,22	78,89	82,22	85,56	76,67	73,33
ⵙ	87,78	90,00	84,44	83,33	86,67	86,67	85,56	85,56
ⵉ	87,78	87,78	86,67	86,67	80,00	82,22	76,67	75,56
ⵊ	77,78	83,33	74,44	73,33	75,56	81,11	72,22	71,11
ⵛ	85,56	88,89	85,56	87,78	83,33	86,67	83,33	85,56
ⵍ	87,78	87,78	84,44	83,33	87,78	87,78	84,44	83,33
ⵎ	80,00	80,00	80,00	78,89	78,89	78,89	78,89	77,78
ⵏ	81,11	86,67	76,67	74,44	80,00	85,56	75,56	73,33
ⵇ	81,11	86,67	78,89	78,89	78,89	84,44	76,67	76,67
ⵔ	85,56	88,89	81,11	78,89	82,22	85,56	77,78	75,56
ⵕ	88,89	90,00	88,89	87,78	85,56	86,67	85,56	84,44
ⵚ	81,11	84,44	77,78	76,67	78,89	82,22	75,56	74,44
ⵜ	83,33	83,33	81,11	78,89	80,00	80,00	77,78	77,78
ⵟ	77,78	81,11	77,78	77,78	75,56	78,89	75,56	75,56
ⵡ	83,33	83,33	80,00	78,89	82,22	82,22	78,89	77,78
ⵅ	81,11	85,56	76,67	74,44	78,89	83,33	74,44	72,22
ⵢ	83,33	85,56	80,00	81,11	83,33	85,56	80,00	78,89
ⵥ	85,56	90,00	83,33	83,33	82,22	86,67	80,00	80,00
ⵦ	83,33	86,67	82,22	82,22	81,11	84,44	80,00	80,00

Table 6: The Amazigh alphabets recognition rates for 3-5 HMMs and different GMMs

3 HMMs			
8 GMMs	16 GMMs	32 GMMs	64 GMMs
82,96	85,76	80,54	79,76
5 HMMs			
8 GMMs	16 GMMs	32 GMMs	64 GMMs
80,67	83,47	78,25	77,47

Table 7: System overall recognition rate

7. **Conclusions**

This paper presents our results for the Amazigh letters speech recognition via an Interactive Voice Response server (IVR) based Moroccan Amazigh language. In our approach to construct the VoIP network, we use Asterisk server as backbone an Amazigh speech recognition system. The presented system recognizes the 33 alphabets using the acoustic model based on HMMs and GMMs. It has been observed from the performed experiments that the accuracy of the proposed system is 85.76%. The best ASR parameterization for the system is 3 HMM and 16 GMMs.

References

Asterisk. (2015) IVR. Retrieved January from: http://www.asterisk.org,.

Shah, K., Ghrera, S. P., Thaker, A. (2012). A novel approach for security issues in VoIP networks in Virtualization with IVR. arXiv preprint arXiv:1206.1748.

Anwar, Z., Yurcik, W., Johnson, R. E., Hafiz, M., Campbell, R. H. (2006). Multiple design patterns for voice over IP (VoIP) security. In Performance, Computing, and Communications Conference. IPCCC 2006. pp. 8.

Rafique, M. Z., Akbar, M. A., & Farooq, M. (2009). Evaluating DoS attacks against SIP-based VoIP systems. In *Global Telecommunications Conference, GLOBECOM 2009.* pp. 1-6.

Basu, J., Bepari, M. S., Roy, R., & Khan, S. (2013). Real time challenges to handle the telephonic speech recognition system. In *Proceedings of the Fourth International Conference on Signal and Image Processing*. pp. 395-408. Springer, India.

Aust, H., Oerder, M., Seide, F., et Steinbiss, V. (1995). "The Philips automatic train timetable information system," Speech Communication. 17(3):249-262.

Bhat, C., Mithun, B. S., Saxena, V., Kulkarni, V., et Kopparapu, S. (2013). Deploying usable speech enabled ivr systems for mass use, Human Computer Interactions (ICHCI).pp. 1-5.

Satori, H., ElHaoussi, F. (2014). Investigation Amazigh speech recognition using CMU tools. *International Journal of Speech Technology*. 17(3): 235-243.

Hamidi, M., Satori, H., et Satori, K. (2016). Amazigh digits speech recognition on IVR server. Advances in Information Technology: Theory and Application. Vol. 1, no 1.

Spencer, M., Allison, M., & Rhodes, C. (2003). The asterisk handbook. *Asterisk Documentation Team*.

Madsen, L., Van Meggelen, J., & Bryant, R. (2011). *Asterisk: The definitive guide*. O'Reilly Media, Inc.

Handley, M. (1999). SIP: session initiation protocol.

Schulzrinne, H. (1996). RTP: A Transport Protocol for Real-Time Applications, Internet Engineering Task Force, Audio-Video Transport w orking Group. RFC1889.

Karapantazis, S., and Pavlidou, F. N. (2009). VoIP: A comprehensive survey on a promising technology", Computer Networks. 53(12):2050-2090.

Huang, X., Acero, A., Hon, H. w ., & Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development.* Vol. 95. Upper Saddle River: Prentice hall PTR.

Beal, M. J., Ghahramani, Z., & Rasmussen, C. E. (2002). The infinite hidden Markov model. *In Advances in neural information processing systems.* pp. 577-584.

TICAM'18, p. 179 - 189

Pathological voice detection using automatic speech recognition based on Amazigh language

Ouissam Zealouk [1,2], Hassan Satori [1,2]
Mohamed Hamidi [1,2], Khalid Satori [1]
[1] Laboratory Computer Science, Image processing and Numerical Analysis,
Faculty of Sciences Dhar Mahraz, Sidi Mohammed Ben Abbdallah University,
Fés, Morocco
[2] Department of Mathematics and Computer Science Faculty Polydisciplinary of Nador, Mohammed Premier University, 300, Selouane 62700 Nador, Morocco.
{ouissam.zealouk,mohamed.hamidi.5,khalidsatori}@gmail.com
hsatori@yahoo.com

Abstract

In the past few years, research on automatic systems to assess voice disorders has received appreciable attention due to its objectivity and noninvasive nature. The work presented in this paper aims to build an automatic speech recognition system based on Sphinx4 that permit to detect persons with voice disorders. This research project is carried out using Amazigh language in order to differentiate between normal and pathological voices. The performance in our system was measured using combinations of HMM 5-states with 8 Gaussian mixture distributions. The obtained results are very satisfying, with a vast difference in recognition rates of normal and pathological speakers.

Keywords: Automatic speech recognition system, Voice disorders, Amazigh language, Hidden Markov Model, Sphinx4.

1. Introduction

Automatic speech recognition (ASR) systems play a vital role in the human-machine interaction. ASR is the process by which a computer converts a speech signal into a sequence of words, also called Speech-to-text systems, through using matching techniques to compare a speech waveform to a set of samples, usually composed of words and phonemes. These systems based on a conventional Hidden Markov Model (HMM), mostly utilize phonemes as basic linguistic units and cepstral features as acoustic observation.

ASR has a vast field of applications, e.g. command recognition, dictation, interactive voice response, learning foreign languages and helping disabled people to interact with society. It is a very promising technology that makes life easier (Halton *et al.*, 2006; Satori *et al.*, 2009). The world of automatic recognition speech has become very flexible and developed thanks to the progress of tools and their availability, which was allowed researchers in recent years to work on detecting dysphonic patients and evaluate their voice (Manfredi, 2009). In our previous work, we employed the ASR technology to develop a system which differentiates between smokers and non-smokers voice using the Amazigh digits speech based on the Mel

frequency spectral coefficients (MFCCs) to determine the voices' features (Satori *et al.*, 2017). Authors of (Maier *et al.*, 2009) have developed an ASR system to evaluate speech and voice disorders. They made the system available on the internet, where a patient can access easily and read text or pictures' name to measure his performance. The results obtained show that the automatic examination and judgment of experts are similar.

In another study, the objective of (Muhammad *et al.*, 2011) was detecting voice disorders based on six different cases. The aim of their work is the classification of the type and severity of voice pathologies using Arabic automatic speech recognition (ASR). The authors of (Godino-Llorente *et al.*, 2006) proposed a detection system of pathological voice by means of Gaussian mixture models and short-term Mel cepstral vectors parameters achieved by framing energy together with first derivatives. The authors of (Wiśniewski *et al.*, 2007) also presented automatic speech recognition system using Hidden Markov Model technique to detect the speech disorders. This system achieved a success rate of approximately 70%. For the purpose of voice pathology evaluation, various databases have been generally used by the researchers, among them Saarbruecken Voice Database and Arabic Voice Pathology Database (Muhammad *et al.*, 2017; w oldert-Jokisz, 2007) these databases are commonly used by the scientific community.

The work presented in this paper aims to develop an automatic speech recognition system capable of analyzing voice speech and detecting whether an Amazigh speaker has a pathological voice or not. The rest of this paper is organized as follows. Section 2 gives an overview of ASR. Section 3 presents a brief description of the Amazigh language. Section 4 emphasizes on the description of Hidden Markov Model. Section 5 shows the technology and the method used in this work. The experimental results are presented in Section 6. Conclusions are drawn in Section 7.

2. Automatic speech recognition

2.1 Basic Principle

Automatic Speech Recognition (ASR) is a special case of pattern recognition (Gaikwad, 2010). This recognition operates in two phases: Training and testing. The process of extraction of features relevant for classification is common in both phases. During the training phase, each reference is learned from spoken examples and stored either in the form of templates obtained by some averaging method or models that characterize the statistical properties of pattern. During the testing or recognition phase, the feature of test pattern (test speech data) is matched with the trained model of each and every class.

The speech recognition system uses acoustic and language models to find the word recognition rate of the speech input that is the number of correctly recognized words. Acoustic modelling plays a very important role in improving the accuracy of the ASR systems. For the given acoustic observation A, the goal of speech recognition is to find the most probable word sequence $\widehat{M}\,\widehat{M}$, that maximize the posterior probability P (M/A), which is written as:

$$\widehat{M} = \underset{M}{\operatorname{argmax}} \; P(A/M)P(M)$$

where A represents the acoustic feature of the word sequence M, P(M) is the language model. The language model contains a set of rules for a language that is used as the primary context for recognizing words. It is important in this process because it helps to reduce the search space and resolve acoustic ambiguity (Huang, 2001). The diagram below (Figure 1) shows a representation of speech recognition system in simple equations, which contains front end unit, model unit, language model unit, and search unit. The recognition process.

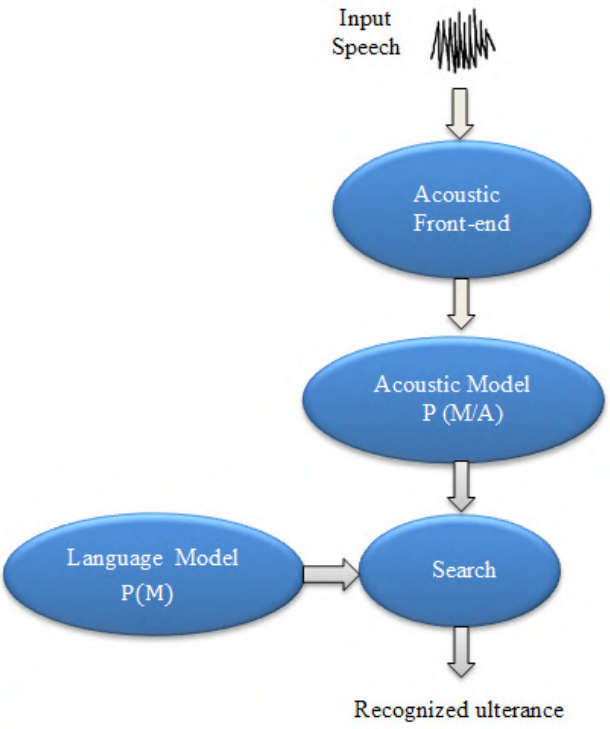

Figure 1: Basic model of speech recognition

2.2. Speech recognition techniques

The goal of speech recognition is for a machine to be able to "hear," understand," and "act upon" spoken information. The earliest speech recognition systems were first attempted in the early 1950s, at Bell Laboratories, to recognize isolated digit for a single speaker (Klevans, 1997). The goal of automatic speaker recognition is to analyze, extract, characterize and recognize information about the speaker identity. The speaker recognition system may be viewed as working in a four stages:

a. Analysis
b. Feature extraction
c. Modeling
d. Testing

A. Speech analysis technique

In speech analysis technique, speech data consist of different types of information that show speaker identity. This includes speaker-specific information due to vocal tract, behaviour characteristic and excitation source. The speech signal is produced from the vocal tract system. The physical construction and dimension of vocal tract as well as the excitation source are unique for each speaker. This information is embedded into the speech signal during speech production, and can be used for speaker recognition. To obtain good representation of these speaker features, speech data needs to be analysed and tested. The speech analysis stage deals with the selection of suitable frame size and frame shift for segmenting the speech signal for further analysis and feature extraction.

B. Feature extraction technique

Feature extraction is the essential part of speech recognition, since it plays an important role in distinguishing one utterance from other [14]. The utterance can be extracted using a vast range of feature extraction techniques suggested and successfully utilized for speech recognition task. But extracted feature should meet some criteria while dealing with the speech signal such as:

1. Easy to calculate extracted speech features.
2. It should not be amenable to mimicry.
3. It should show little change from one speaking environment to another.
4. It should be steady over time.
5. It should occur ordinarily and naturally in speech.

3. Amazigh language

The Amazigh language or Tamazight is spoken in a vast geographical area of North Africa, from the Canary Islands to the Siwa Oasis in the North, and from the Mediterranean coast to Niger, Mali and Burkina Faso in the South. Historically, the Amazigh language has been autochthonous and was exclusively reserved for familial and informal domains (Boukous, 1995).

In Morocco, the Amazigh language is spoken by some 28% of the population, grouped in three main regional varieties, depending on the area and the communities: Tarifit spoken in northern Morocco, Tamazight in central Morocco and South-East, and Tachelhit spoken in southern Morocco (Ouakrim, 1995; Chaker, 1984).

Since 2003, Tifinaghe-IRCAM has become the official graphic system for writing Amazigh in Morocco. This system contains (Ataa Allah and Boulaknadel, 2012):

- 27 consonants including the labials (ⵀ , ⵃ, ⴼ), the alveolars (ⵟ , ⵙ, ⵣ, ⵚ), the dentals (ⵜ , ⵏ , ⴹ , ⴷ , ⵍ , ⵔ , ⵥ , ⵝ), the palatals (ⵛ , ⵊ), the velar (ⴽ, ⵅ), the labiovelars (ⴽ’ , ⵅ’), the uvulars (ⵇ , ⵅ , ⵖ), the pharyngeal (ⵄ , ⵃ) and the laryngeal (ⴰ);
- 2 semi-consonants: ⵢ and ⵡ ;

- 4 vowels: three full vowels: ⵄ, ⵅ, ⵓ, and neutral vowel (or schwa) ⴻ, which has a rather special status in Amazigh phonology.

The allowed syllables in Amazigh language are: V, CV, VC, CVC, C, CC and CCC, where V indicates a vowel while C indicates a consonant (Ridouane, 2003). In this study, we are interested in speech recognition of Amazigh digits. Table 1 shows a description of the 10 first Amazigh digits used in our system.

Digits	English transcription	Arabic transcription	Tifinaghe transcription	Syllables
0	AMYA	أميا	ⵄ ⵎ ⵅ ⵄ	VC-CV
1	YEN	يان	ⵅ ⵄ ⵏ	CVC
2	SIN	سين	ⵙⵅ ⵏ	CVC
3	KRAD	كراض	ⴽ ⵕⵄ ⴹ	VC-CVC
4	KOZ	كوز	ⴽ ⴽ ⵓ ⵣ	CVC
5	SMMUS	سموس	ⵙⵎ ⵎ ⵓ ⵙ	CC-CVC
6	SDES	سضيس	ⵙⴹ ⵅ ⵚ	CCVC
7	SA	سا	ⵙⵄ	CV
8	TAM	تام	ⵜ ⵄ ⵎ	CVC
9	TZA	تزا	ⵜ ⵣ ⵄ	CC-CV

Table 1: The ten first digits with transcription in English, Arabic and Amazigh letters and their syllables.

4. Hidden Markov Model

The Hidden Markov Model (HMM) is a popular statistical tool for modeling a wide range of time series data. It provides efficient algorithms for state and parameter estimation, and it automatically performs dynamic time warping of signals that are locally stretched. Hidden Markov models are based on the well-known chains from probability theory that can be used to model a sequence of events in time. The Markov chain is deterministically an observable event. The most likely word with the largest probability is produced as the result of the given speech waveform. A natural extension of the Markov chain is the Hidden Markov Model, where the internal states are hidden and any state produces observable symbols or observable evidences (Young *et al.*, 2002). Mathematically Hidden Markov Model contains five elements.

1. Internal States: These states are hidden and give the flexibility to model different applications. Although they are hidden, usually there is some kind of relation between the physical significance to hidden states.
2. Output: O = {$O_1 O_1$, O_2, O_2, O_3 ,O_3 ,…, O_n, O_n} an output observation alphabet.
3. Transition Probability Distribution: A = a_{ij} is a matrix. The matrix defines what the probability to transit from one state to another is.
4. Output Observation: Probability Distribution B = b_i (k) is the probability of generating observation symbol o (k) while entering to state i.
5. The initial state distribution ($\pi = \{\pi_i\}$) is the distribution of states before jumping into any state.

Here the three symbols A, B and π, represent the probability distributions. They are usually written in HMM as a compact form denoted by lambda as λ = (A, B, π) (Ganesh and Sunil, 2014).

The basic HMM model used in this work is 5-states HMMs architecture for each Amazigh phoneme: three emitting sates and two non-emitting ones as entry and exit which join models of HMM units together in the ASR engine, as shown in Figure 2. Each emitting state consists of Gaussian mixtures trained on 13 dimensional coefficients MFCC and their delta and delta vectors, which are extracted from the signal.

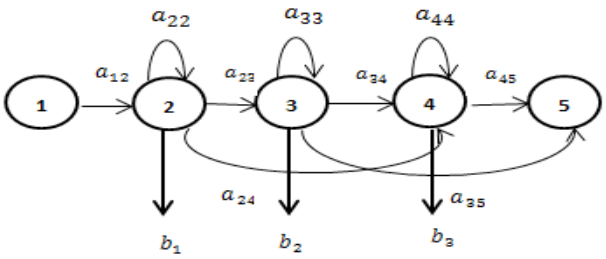

Figure 2: Hidden Markov Model (HMM)—5-states

5. Technology and method

5.1 Sphinx -4

The Sphinx4 is a system of speech recognition designed by Carnegie Mellon University (CMU), Sun Microsystems laboratories, Mitsubishi Electric Research Labs, and Hewlett-Packard's Cambridge Research Lab. It is developed completely in the Java TM programming language. Sphinx4 uses newer search strategies, is universal in its acceptance of various kinds of grammars and language models, types of acoustic models and feature streams[1].

1 http://cmusphinx.sourceforge.net

5.2 SphinxtTrain

SphinxTrain is CMU tool used for acoustic models' development. It is a set of programs and documentation to realize constructing acoustic models for several languages (Satori *et al.*, 2007).

5.3 Speech database preparation

This phase consists of recording the speech signal. Firstly, we used a desktop microphone in clean environment and wavesurfer tool, while keeping a distance of approximately 5-10 cm between mouth of the speaker and the microphone. The sampling rate used for recording is 16 kHz, with 16 bits resolution for more details on the corpus. Technical parameters are given in Table 2. The database used in our system includes recording of 22 Amazigh speakers aged between 26 and 50 years. This database is divided in two categories: the first one consists of 20 normal persons and the second contains 2 speakers having vocal fold disorders (the voice recording with disorders patients has been approved by the ethical committee of Oujda University Hospital). The method followed to record voice is asking those speakers to pronounce the ten first Amazigh digits (ten times for each digit) sequentially. Audio recordings for each speaker was saved in ten ".wav" files, every ".wav" file includes ten repetition of one number. Then, we divided each file into ten wav files. Thus, this corpus consists of 2200 tokens.

Parameter	**value**
Sampling rate	16 kHz
Number of bits	16 bits
Channels	1 (Mono)
Audio data file format	.wav
Corpus	10 Amazigh-digits
Accent	Moroccan Tarifit Berber

Table 2: System parameters.

5.4 Acoustic model

The acoustic model consists of the sub-words which are called phonemes that collectively form the word. It allows converting the pronounced words into phonemes and from phonemes to words that are a statistically possible representation of the acoustic image for the voice signal. During the learning and training Phase, each acoustic unit or phoneme is represented

by a statistical model describing the distribution of data. The speech signal is transformed into a series of features vectors including MFCC coefficients (Mel-Frequency Cepstral coefficients) (Varela *et al.*, 2003).

The Sphinxbase and Sphinxtrain are used to generate the acoustic models. Every recording in the training corpus is transformed into a sequence of feature vectors. The front end provided by Sphinxtrain computes an ensemble of features files for each recording. In this work, the acoustic model was generated using speech signal from the Amazigh digits training database.

5.5 Dictionary

The dictionary provides pronunciations for each existing word in the Language Model and it includes the words we want to train followed by their pronunciation, which divides words into sequences of sub-word units. Our dictionary includes the symbolic representations of the first ten Amazigh digits. The alternative transcripts marked with parentheses as (1) stand for the second pronunciation. The pronunciation dictionary is considered an intermediary between the language model and acoustic model.

5.6 Language model

The language model is defined in three kinds: the simplest that is used for isolated word recognition, the second who is for applications based on command and control, and the last that is a set of n-gram grammars used for free speech form. Each word in the language model should be in the dictionary. In our work, we used a grammar file that includes the ten first Amazigh digits which are shown in Figure 3.

```
#JSGF V1.0;
/**
 * JSGF Grammar for amdigits example
*/
grammar amdigits;

public <greet> = (Amya | Yen | Sin | Krad | Koz | Smmus | Sdes | Sa |
Tam | Tza);
```

Figure 3: The grammars file of Amazigh digits.

6. Experimental Results

The two experiments, in this work, were conducted on a connected phoneme task constituting isolated ten first Amazigh digits. Each phoneme was modeled by a five state HMM. The number of mixture in the model of each state was 8. The first experiment concerns the training and testing of the system with the normal speakers (18 training 2 tests). The second experiment is about testing the system performance with the speakers, who have pathological voices (training by using the voice of 18 normal people and testing by 2 pathological voices).

Figure 4 shows a comparison of the system's performance for the ten first Amazigh digits using 8 GMM (Gaussian mixture models) with 5 HMM for normal voice and voice disorders.

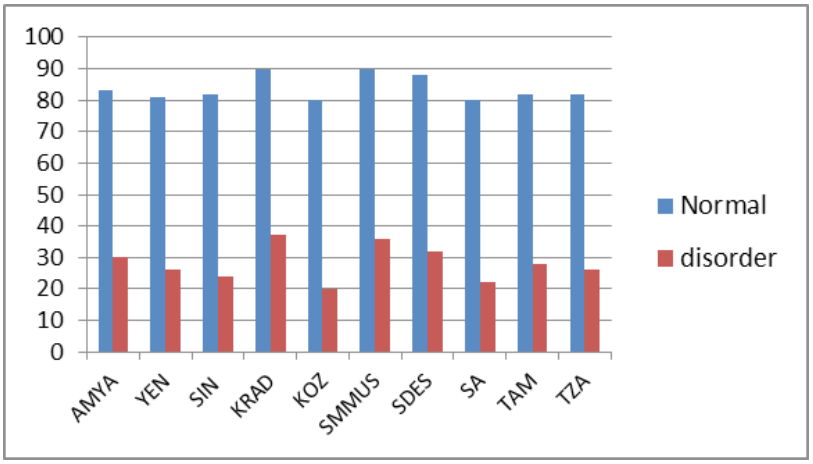

Figure 4: Recognition accuracy (%) of normal and pathological voices

From Figure 4, it is clear that the recognition accuracy obtained for Amazigh digits spoken by normal speakers was very high compared to pathological speakers. There was a significant loss of accuracy on speech recognition for voice disorder samples. The recognition rate by the normal speakers was above or equal to 80%, while the highest rate for pathological speakers reached 35%. This difference between recognition rate of normal and speakers suffering from voice disorders is due to the speech signal of a subject with disorders containing lower amplitude than the speech signal of a normal subject (Zulfiqar *et al.*, 2016), in addition to the impairment of mucosal vibration. Based on the results obtained from the experiences we can see that our system is able to distinguish between the normal and pathological voices.

7. Conclusion

In this study, we suggested an approach to detect speakers who have voice disorders based on automatic speech recognition system, this designed system was developed using open source CMU Sphinx 4 depend on the ten first Amazigh digits. We believe that this is the first study that tries to evaluate the accuracy of ASR in Amazigh speech for people with pathological voices. In our future work we will record larges number of pathological speakers and investigate the performance of the proposed system in continuous speech to analyze different kinds of vocal tract disorders.

References

Ataa Allah, F., Boulaknadel, S. (2012). Natural language processing for Amazigh language: Challenges and future directions. Proceeding of Language Technology for Normalisation of Less-Resourced Languages.

Boukous, A. (1995). Société, langues et cultures au Maroc: Enjeux symboliques. Najah El Jadida, Casablanca, Maroc.

Chaker, S. (1984). Textes en linguistique berbère: introduction au domaine berbère. Paris: Ed. du C.N.R.S.

Desai, N., Dhameliya, K., Desai V. (2013). Feature extraction and classification techniques for speech recognition: A review. International Journal of Emerging Technology and Advanced Engineering. 3(12):367-37.

Gaikwad, S. K., Gawali, B. w., Yannawar, P. (2010). A review on speech recognition technique. International Journal of Computer Applications. 10(3):16-24.

Ganesh, S. P., Sunil, S. M. (2014). Realization of Hidden Markov Model for English Digit Recognition. IJCA. vol. 98, n°17.

Godino-Llorente, J., Gomez-Vilda, P., Blanco-Velasco, M. (2006). Dimensionally reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters. IEEE Trans. Biomed. Eng. 53 (10):1943-1953.

Halton, M., Cerisara, C., Fohr, D., Laprie, Y., Smaili, K. (2006). Reconnaissance automatiue de la parole du signal a son interpretation, Monographies and books, Oxford.

Huang, X., Acero, A., Hon, H., Foreword, B. (2001). Spoken language processing: A guide to theory, algorithm, and system development. Prentice Hall PTR.

Klevans, R. L., Rodman, R. D. (1997). Voice recognition. Artech House, Inc.

Maier, A. Haderlein, T., Eysholdt, U. (2009). PEAKS–A system for the automatic evaluation of voice and speech disorders. Speech Communication. 51(5):425-43.

Manfredi, M. K. (2009). New trends in voice pathology detection and classification. Biomedical Signal Processing and Control, Editorial. 4, pp. 171-172. doi:10.1016/j. bspc..07.001.

Muhammad, G., Alsulaiman, M., Ali, Z., Mesallam, T., Farahat, M. . Malki, K., Al-nasheri, A. (2017). Bencherif MA Voice pathology detection using interlaced derivative pattern on glottal source excitation. Biomed Signal Process Control, Vol. 31, pp. 156–164.

Muhammad, G., Mesallam, T. A., Malki, K. H., Farahat, M. Alsulaiman, M., Bukhari, M. (2011). Formant analysis in dysphonic patients and automatic Arabic digit speech recognition. Biomedical engineering online. 10(41):1-12.

Ouakrim, O. (1995) Fonética y fonología del Bereber. Survey: University of Autònoma de Barcelona.

Ridouane, R. (2003). Suites de consonnes en berbère: phonétique et phonologie. Doctoral Dissertation, Université de la Sorbonne nouvelle-Paris III.

Satori, H., Harti, M., Chenfour, N. (2007). Arabic speech recognition system using cmu-sphinx4. arXiv preprint arXiv:0704.2201.

Satori, H., Hiyassat, H., Harti, M., Chenfour, N. (2009). Investigation Arabic speech recognition using CMU sphinx system, The International Arab Journal of Information Technology. Vol. 6, no. 2.

Satori, H., Zealouk, O., Satori, K. (2017). Voice comparison between smokers and non-smokers using HMM speech recognition system. International Journal of Speech Technology. 20(4):771-777.

Varela, A., Cuayáhuitl, H., Nolazco-Flores, J.A. (2003). Creating a Mexican Spanish Version of the CMU Sphinx-III Speech Recognition System. Springer. Vol. 2905.

Wiśniewski, M., Kuniszyk-Jóźkowiak, W., Smołka, E., Suszyński, W. (2007). Automatic Detection of Disorders in a Continuous Speech with the Hidden Markov Models Approach. In Computer Recognition Systems 2, Vol. 45, pp. 447-453.

w oldert-Jokisz, B. (2007). Saarbruecken Voice Database.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., w oodland, P. (2002). HTK Book, http://htk.eng.cam.ac.uk.

Zulfiqar, A., Alsulaiman, M., Elamvazuthi, I. (2016). Voice pathology detection based on the modified voice contour and SVM. Biologically Inspired Cognitive Architectures. Vol. 15, pp. 10-18.

TICAM'18, p. 191 - 203

Contribution au développement de corpus et système de reconnaissance vocal pour la langue amazighe

Safâa EL OUAHABI, Mohamed ATOUNTI, Mohamed BELLOUKI
Laboratoire MASI, Faculté Pluridisciplinaire de Nador, Université Mohamed Premier Oujda
safaa.elouahabi@gmail.com atounti@hotmail.fr mbellouki@hotmail.com

Résumé

Avec la diffusion de la langue amazighe sur le web et la disponibilité des moyens de manipulation informatique de cette langue, des travaux de recherche ont abordé des problématiques plus variées pour qu'elle puisse rejoindre ses consœurs dans le domaine des nouvelles technologies de l'information et de la communication. Dans ce contexte, de nombreux chercheurs au niveau national ont projeté leurs attentions sur le développement de toutes les applications qui peuvent améliorer la situation actuelle de la langue amazighe. Ce travail a pour objectif principal la contribution au traitement automatique de la parole amazighe, dans lequel nous avons étudié et réalisé un système de reconnaissance automatique de la parole en utilisant un environnement basé entièrement sur la langue amazighe. Dans ce cadre, nous avons d'abord construit une base de parole composée de 573 mots isolés prononcés par 50 locuteurs. 150 mots sélectionnés de cette base ont été utilisé et évalué en développant un système de reconnaissance de la parole. Les résultats de la reconnaissance montrent que le système est indépendant et sa performance atteint un taux de 70%. Ces résultats sont très encourageants et nous favorisent de recueillir un corpus oral plus riche et varié composé de mots amazighes dédiés à l'apprentissage de cette langue.

Mots clés : *Langue amazighe, traitement automatique de la parole, base de parole amazighe, mots isolés.*

1. Introduction

De nos jours, les ordinateurs sont largement utilisés pour communiquer en utilisant le texte et la parole. Des outils de traitement de texte, des dictionnaires électroniques, des services en ligne, ou des systèmes plus avancés tels que ceux dédiés à la dictée ou à la synthèse vocale, sont disponibles pour un petit nombre de langues parmi les 7000 parlées dans le monde. Dans le cadre de ce travail, nous nous concentrons sur la langue amazighe qui a peu de ressources informatiques. Le but cependant est clair, c'est de faire de l'amazighe une langue écrite et parlée équipée et dotée de ses références, au service de tous les locuteurs et utilisateurs, une langue accessible qui répond aux besoins des utilisateurs dans les situations de communication requises dans notre vie moderne.

Fondamentalement, le corpus de la parole est nécessaire pour former tous les systèmes de la reconnaissance de parole, La rareté des données d'apprentissage parlé et écrit est l'un des principaux problèmes auxquels les chercheurs sont confrontés. Il y a eu plusieurs travaux et beaucoup d'efforts dans le développement du corpus de l'anglais et d'autres langues majeures (Robinson *et al.*, 1995 ; Hoge *et al.*, 1997 ; Al-Sulaiti et Atwell, 2006 ; Abushariah *et al.*, 2010). Dans le cadre de notre recherche, nous nous sommes principalement concentrés sur la collecte d'un corpus de parole amazighe à mots isolés et le développement d'un système de reconnaissance de parole. La motivation derrière la création de notre corpus de mots isolés est de collecter des enregistrements de la langue amazighe, adaptés aux différents services tels que les applications d'apprentissage, de la dictée, de la synthèse vocale et de la reconnaissance automatique de la parole.

Les chercheurs ont ciblé et projeté leurs attentions au développement des systèmes de reconnaissance vocale automatique continue, discrète, de petit, moyen et grand vocabulaire pour différentes langues (Jitendra et Sanjay, 2014 ; Vrinda et Chander, 2013 ; Kimutai *et al.*, 2013 ; Ananthi et Dhanalakshmi, 2013 ; Kumar *et al.*, 2012 ; Charansing *et al.*, 2012 ; Ghai et Singh, 2012; Sameti *et al.*, 2011). Le paradigme d'apprentissage de n'importe quel système nécessite des fichiers de voix audio avec leur texte transcrits afin d'estimer les paramètres des modèles. Ainsi, les bases de données vocales sont des ressources critiques dans le développement d'un système de reconnaissance de la parole ainsi que leurs caractéristiques, comme le nombre d'heures d'enregistrement, le nombre de locuteurs, etc. De plus, le critère de l'indépendance du système du locuteur implique des ressources d'un nombre important. La variabilité de la parole est un facteur important et même critique en modélisant complètement et précisément les différentes prononciations possibles de chaque personne. Cela peut, à son tour, être réalisé en utilisant des enregistrements d'un grand nombre pour un même locuteur.

Nous présentons, dans cet article, les étapes de la collecte des voix de notre base. Puis, nous décrivons les différentes phases de construction des fichiers audio pour la base de voix crée dans le cadre de cette recherche. Nous exposons aussi les caractéristiques et les détails statistiques et techniques de la base.

Nous passons ensuite à l'étape de l'adaptation du corpus de parole au système de reconnaissance de la parole. Cette étape nécessite de traiter le corpus de parole pour se conformer aux normes du système de reconnaissance CMU Sphinx.

2. Travaux antérieurs sur la reconnaissance de la parole amazighe

Il y a beaucoup d'efforts fournis pour développer un système de reconnaissance de parole pour la langue Amazighe au Maroc. Certains travaux sont effectués dans cette direction pour les chiffres et les alphabets Dans cette section, nous citons des travaux publiés qui touchent la reconnaissance de la parole amazighe.

El Ghazi *et al.* (2011) ont exploité les règles de construction des chiffres amazighe pour construire un système de reconnaissance automatique de la parole. Ce système est basé sur la synthèse de chiffres enchaînés à partir d'un nombre isolé de 1 à 10. Il apparaît comme un outil important pour minimiser le corpus d'entraînement, il évite le chevauchement entre différentes prononciations. Ce système permet également d'étendre la reconnaissance

automatique du dialecte amazighe. Les résultats obtenus sont satisfaisants par rapport à la taille de la base de données d'apprentissage. Par conséquent, leur système constitue une première partie d'un système de sécurité.

El Ghazi *et al.* (2014) présentent un système de reconnaissance automatique de la parole amazighe basé sur la transcription en alphabet Tifinaghe reconnue par l'Institut Royal de la Culture Amazighe (IRCAM). Ils ont utilisé le modèle de Markov caché et l'ont comparé avec la méthode de programmation dynamique. Leur travail donne une idée sur la phonétique utilisée pour la reconnaissance de cette langue. En comparaison, avec la programmation dynamique, Les résultats obtenus par le modèle de Markov caché (HMM, Hidden Markov Model) sont très satisfaisants malgré la limitation du nombre de locuteurs et de la taille de la base de données. Ceci montre l'importance de modélisation stochastique et probabiliste dans le domaine de la reconnaissance.

Telmem *et al.* (2014) ont développé un système de reconnaissance automatique de la parole pour la langue amazighe. Ce système est basé sur les modèles de Markov cachés avec des mélanges gaussiens pour générer des modèles acoustiques. Ils ont utilisé l'outil CMU Sphinx-4. La taille de la base de voix pour ce travail de recherche est de 11220 mots et le système a atteint un taux de reconnaissance de 90%. Le système produit dans le cadre de ce travail n'est pas optimal mais ses performances sont satisfaisantes par rapport aux autres systèmes développés.

Satori *et al.* (2014) ont étudié l'indépendance du système de reconnaissance de la parole en utilisant une base de voix « alphadigits » correspondant aux chiffres et lettres prononcés par des locuteurs d'origines amazighes. Ce système a été mis en œuvre en utilisant CMU Sphinx. Il est basé sur les Modèles de Markov cachés. Ce travail comprend la création de la base de voix alphadigits, qui se compose des chiffres et lettres amazighes utilisés dans la phase d'apprentissage et de test du système. Les résultats de la reconnaissance montrent que leur système est indépendant et sa performance est satisfaisante.

Dans la contribution d'(Abdenbi *et al.*, 2018), les auteurs proposent un système de reconnaissance automatique des mots isolés de la langue amazighe basé sur les transformations orthogonales paramétrables. Ils ont pu atteindre un taux de reconnaissance plus élevé qui tend vers les 96%. Cependant, leur travail est une première initiative pour la réalisation d'un système de reconnaissance de la parole amazighe assurant l'apprentissage de la prononciation, qui suscite l'intérêt de recueillir un corpus oral riche et varié composé de mots amazighes dédiés à l'apprentissage de la langue.

3. Bases théoriques

3.1. Les Modèles de Markov Cachés

Les Modèles de Markov Cachés (MMC), en anglais Hidden Markov Model (HMM), sont des automates doublement stochastiques permettant de modéliser des données séquentielles dans de nombreux domaines, tels que le traitement automatique du langage naturel, l'intelligence artificielle, la reconnaissance de formes, la biologie, l'indexation de documents…

Le processus stochastique (ou processus aléatoire) est un processus qui modélise l'état du système et lui-même observé au moyen d'une émission.

Un Modèle Markov caché est un processus stochastique (ou séquence de variables aléatoires) qui suit la loi d'une chaîne de Markov. Ce processus peut parfois être expliqué par un autre processus caché qui suit la loi d'une chaîne de Markov. Ce second processus est dit caché car le premier, celui qu'il explique, est le seul observé.

- Un processus stochastique est un ensemble de variables aléatoires définies sur un espace de probabilités.
- Une variable aléatoire (d'observations) est une fonction mesurable définie sur un espace de probabilités avec des valeurs dans un ensemble de réalisations (appelé parfois ensemble d'états).

La modélisation d'un Modèle de Markov Caché est largement détaillée et illustrée mathématiquement (Rabiner, 1989), leur principe de modélisation est de mesurer une variable aléatoire X(t) et de chercher à modéliser le processus par un modèle stochastique (MMC) défini par M(A,B,π) comme illustré sur la figure 1, telles que :

- *A* est la matrice des probabilités de transition entre les états. La probabilité de transition d'un état *Si* vers un autre état (ou lui-même) est $A = \{a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i)\}$ avec $1 \leq i, j \leq N$, N indique le nombre d'états de la chaîne de Markov $S = \{S_1, S_2, ..., S_N\}$, q_t indique l'état à l'instant *t* .
- *B* est la matrice des probabilités d'émission des observations dans chaque état $B = \{b_i(O_k) = P(O_k \mid q_t = S_i)\}$ avec $1 \leq i \leq N$ et $1 \leq k \leq M$
- $O = \{O_1, O_2, ..., O_M\}$ ce sont des symboles émis par les états.
- π est le vecteur des probabilités initiales des états. Pour chaque état S_i la probabilité d'être atteint à partir de l'état initiale q_1 est π_i, $\pi = \{\pi_i\} = \{P(q_1 = S_i)\}$ et $1 \leq is \leq N$.

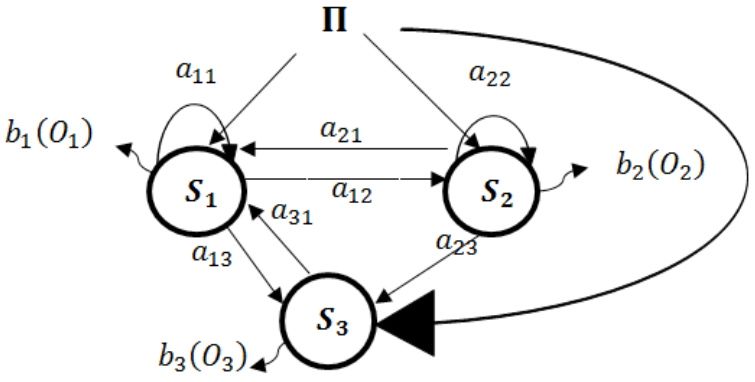

Figure 1 : Représentation d'une chaîne de Markov Caché

La topologie utilisée dans la conception de notre système est le modèle Bakis ou Modèle Gauche-Droite (voir figure 2). C'est un modèle avec contrainte sur les connections entre les états. La transition d'un état ayant un indice bas vers un état ayant un indice haut, c'est-à-dire n'autorise aucune transition d'un état d'indice supérieur vers un autre état d'indice inférieur (pas de retour en arrière). La topologie gauche-droite offre des avantages pour utilisation, elle est convenable à la modélisation des signaux et aussi à la phase de l'apprentissage. Il y a moins de paramètres à estimer. Cette topologie est très utilisée dans les systèmes de la reconnaissance automatique de l'écriture et en reconnaissance automatique de la parole.

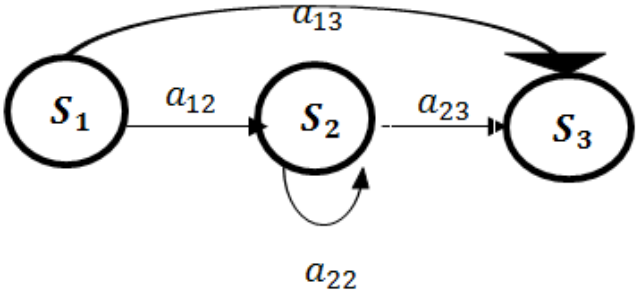

Figure 2 : Modèle de Bakis

4. Le corpus de la parole amazighe

Dans le contexte de la recherche scientifique en langue amazighe, les données recueillies par les chercheurs ne sont pas mis à disposition de l'ensemble de la communauté scientifique. Il existe ainsi, une multitude de données orales et écrites sur la langue amazighe mais leur accès est limité.

En ce qui concerne l'écrit, des bases de données textuelles existent, par exemple la contribution de (Boulaknadel et Ataa Allah, 2013).

Tandis que les corpus oraux sont difficilement accessibles par des personnes extérieures à la recherche locale. Les données sont limitées et dépendent de l'objectif de l'étude et de la connaissance des outils à disposition de chaque chercheur. D'autre part, dans un contexte d'apprentissage statistique les données doivent être disponibles en grande quantité.

Partant de l'idée que tout traitement automatique de la langue amazighe ne peut se faire sans avoir un corpus oral libre et accessible, nous consacrons une grande importance à ce type de travail. Ainsi, nous avons constitué et analysé un corpus composé de 573 mots isolés amazighes.

4.1. Caractéristiques du Corpus

La collecte de données fait généralement partie intégrante du développement de n'importe quel service vocal.

Dans le cadre de notre recherche, nous nous sommes principalement concentrés sur la collecte d'un corpus de parole amazighe, et nous avons développé sa transcription phonétique. Les caractéristiques de notre corpus peuvent être déclinées comme suit :

1) Il contient 573 mots distincts.
2) La transcription phonétique de tous les mots enregistrés.
3) Le corpus peut être utilisé pour la l'apprentissage ainsi que pour tester tout système supportant la langue amazighe.
4) Le corpus est recueilli à partir de locuteurs appartenant à diverses régions parlant tarifit (Région de l'oriental).
5) Le corpus peut être utilisé pour le développement de nombreuses applications basées sur la parole et le texte amazighe.
6) Le corpus peut être adapté aux différents services, tels que les applications d'apprentissage, les systèmes de la dictée et de la synthèse vocale.

4.2. Préparation du corpus

Le corpus vocal est enregistré à l'aide d'un microphone en utilisant l'outil wavesurfer1 au format wav(.wav). Le corpus est constitué de 573 mots amazighes recueillis auprès de 50 locuteurs marocains natifs de Tarifit (25 hommes et 25 femmes) âgés de 18 à 40 ans (le tableau 1 montre la répartition de l'âge et du genre). La fréquence d'échantillonnage de l'enregistrement est de 16 kHz, avec une résolution de 16 bits (Voir tableau 2). Pendant les sessions d'enregistrement, les locuteurs ont été invités à prononcer les mots isolés séquentiellement. Les enregistrements de chaque locuteur ont été enregistrés dans un fichier «.wav». Chaque fichier a été relu pour s'assurer que tous les mots étaient inclus dans le signal enregistré. Par conséquent, les fichiers .wav sont segmentés en plus petits fichiers ayant chacun un seul enregistrement d'un seul mot. Les mots mal prononcés ont été ignorés et seuls les mots corrects sont conservés dans la base.

N°	***Catégorie d'âge***	***Genre***		***Totale***
		H	F	
1	Moins de 30 ans	20	20	40
2	Plus que 30 ans	5	5	10
Totale		25	25	50

Table 1 : Locuteurs selon l'âge et le genre

1 http://sourceforge.net/projects/wavesurfer/

Détails techniques de l'enregistrement	Valeur
Fréquence d'echantillonage (kHz)	16
résolution (bits)	16
Channels	1(Mono)
Format des fichiers audios	.wav
Corpus	573 mots
Accent	Tarifit (Région de l'oriental)

Table 2 : Attributs d'enregistrements

Il a apparu dans la phase de vérification audio que certains fichiers enregistrés étaient bruyants, principalement à cause du bruit au niveau du microphone et des connexions audio et aussi à cause des interférences et des bruits environnementaux. Afin d'éliminer le bruit, nous avons manipulé les fichiers en utilisant le logiciel audacity[2] qui offrent la technique de réduction du bruit (voir figures 3 et 4).

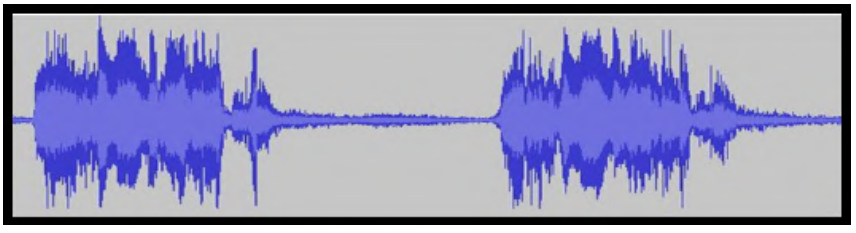

Figure 3 : Signal avant élimination du bruit

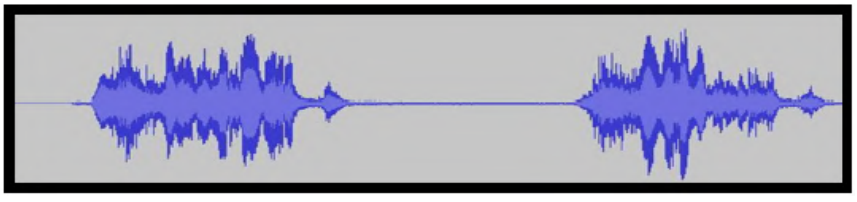

Figure 4 : Signal après élimination du bruit

2 www.audacity.fr

5. Système de reconnaissance de la parole amazighe basé sur CMU Sphinx

Les technologies de reconnaissance vocale particulièrement pour la variation tarifit peut être considérées moins couteuses, en terme de développement, que d'autres liées aux dictionnaires et traitement de texte. Nous argumenterons qu'un système de traitement de la parole peut traiter des ressources adaptées aux différents services, tels que les applications d'apprentissage, les systèmes de la dictée, de la synthèse vocale, et de la traduction vers d'autres langues. Dans le but de notre recherche, nous décidons de regrouper tout d'abord les locuteurs selon le critère de la disponibilité (régions de l'oriental), et élargir en futur l'échantillon de notre étude pour couvrir les différentes variantes de la langue amazighe.

Après avoir construit le corpus, nous avons passé à l'utilisation de ce corpus pour développer un système de reconnaissance automatique de la parole indépendant. Afin d'atteindre ces objectifs, le travail a été divisé en parties distinctes. Chacun cible à atteindre un objectif partiel. Le produit final est la combinaison de toutes les sous-parties. Ces étapes peuvent être résumé comme ci-dessous :

- Conception d'un corpus de parole : développer un corpus qui puisse être lu et enregistré (voir section 4).
- Adaptation du corpus de parole au système de reconnaissance de la parole. Cette étape nécessite de traiter le corpus de parole pour se conformer aux normes du système de reconnaissance CMU Sphinx.
- Apprentissage et test du système de reconnaissance de parole. La tâche principale consiste à étudier le fonctionnement interne du système et à régler tous les paramètres pour produire des résultats satisfaisants.

5.1. Préparation des fichiers audio

Nous avons sélectionné une partie de notre corpus d'environ 150 mots amazighes et nous avons préparé les fichiers acquis par le toolkit CMU-Sphinx.

Fichier de transcription :

Tous les mots prononcés sont transcrits, même le silence ou le bruit devrait être représenté dans le fichier de transcription. La transcription est alors suivie par le nom du fichier sans le chemin comme indiqué sur la figure 5.

```
<s> AMAN </s> (abdoaman1)
<s> AMAN </s> (abdoaman2)
<s> AMAN </s> (abdoaman3)
<s> AMAN </s> (abdoaman4)
<s> AMAN </s> (abdoaman5)
```

Figure 5 : Exemple de fichier de transcription

Ainsi, dans un fichier, les mots prononcés par le même locuteur sont cités de la même manière qu'ils ont été enregistrés, avec une balise de silence (balise de début <s>, balise de fin </ s>), suivie de l'ID du fichier représentant le mot. Ce fichier est connu comme fichier de

transcription et porte l'extension '.transcription'. On prépare deux fichiers, un d'entre eux est utilisé pour entraîner le système et un autre pour les tests.

Dictionnaire phonétique :

Le dictionnaire phonétique sert d'intermédiaire entre le modèle acoustique et le modèle de langage. Le dictionnaire de prononciation est constitué d'enregistrements contenant des mots et des séquences monophones associées (voir figure 6).

```
AZWA  AE Z W AH
BABA  B AH B AH
BABA(2)    B AA B AH
```

Figure 6 : Extrait du dictionnaire phonétique

Dictionnaire filler :

Ce dictionnaire répertorie généralement les événements non vocaux en tant que «mots» et les mappe sur des phonèmes définis par l'utilisateur. Ce dictionnaire doit au moins avoir les entrées :

- <s>: silence de début d'énoncé ;
- <sil>: silence d'intérieur ;
- </ s>: silence de fin.

Notez que les mots <s>, </ s> et <sil> sont traités comme mots spéciaux et sont tenus d'être présents dans le dictionnaire. Au moins l'un d'entre eux doit être mappé sur un phonème appelé «SIL» (voir figure 7). Le phonème SIL est traité d'une manière spéciale et doit être présent. Pour un discours propre, ces événements peuvent en fait être des silences, mais pour un discours bruyant, ils peuvent être le type de bruit de fond le plus général qui règne dans la base de données. D'autres bruits peuvent ensuite être modélisés par des phonèmes définis par le développeur du système.

```
<s> SIL
</s> SIL
<sil> SIL
```

Figure 7 : Dictionnaire filler

Liste des phonèmes:

La liste de phonèmes est une liste de toutes les unités acoustiques pour lesquelles nous voulons former des modèles. La liste de phonèmes doit avoir exactement les mêmes unités utilisées dans le dictionnaire, ni plus ni moins. Chaque phonème doit figurer sur une ligne distincte du fichier, en commençant par la gauche, sans espace supplémentaire après le phonème. Un exemple est montré sur la figure 8.

```
AY
B
CH
D
EH
ER
EY
F
G
```

Figure 8 : Extrait de la liste de phonèmes

5.2. Extraction de caractéristiques du signal

L'extraction de caractéristiques est la première phase de construction du système de reconnaissance de la parole. Il est responsable de convertir le signal vocal en vecteur de caractéristiques afin d'être utilisé pour l'apprentissage et le test du système. Il est responsable de la collecte, de l'annotation et du traitement des données d'entrée. Il extrait les fonctions à lire par le décodeur. L'extraction des caractéristiques implique l'analyse du signal vocal. Les Coefficients Cepstraux Mel-Frequency (MFCC) ont été appliqués pour extraire des caractéristiques du signal vocal. Les paramètres utilisés dans notre système étaient un taux d'échantillonnage de 16 KHz avec un échantillon de 16 Kbit, une fenêtre de Hemming de 25,6 ms avec des cadres consécutifs se chevauchant de 10 ms et des coefficients cepstraux Mel-Frequency (MFCC).

5.3. Création du modèle de langage

Le but principal de notre travail concernant la langue amazighe était de développer un modèle de langage approprié pour la parole continue amazighe. Nous avons utilisé toutes les transcriptions textuelles des fichiers audio de la base de parole que nous avons conçus pour créer un modèle de langage à trois grammes.

5.4. Création du modèle acoustique

Le modèle acoustique est la représentation de la grammaire ou de la syntaxe de la tâche. Les systèmes de reconnaissance de parole à grand vocabulaire utilisent des modèles de Markov cachés de type Bakis (HMM) avec des modèles de mélange gaussiens (GMM) comme distributions de sortie pour modéliser des sous-mots comme unités vocales telles que les phonèmes dépendant du contexte (tri-phones) ou les tied model (tied state). Les HMM modélisent ces unités vocales en utilisant des vecteurs de caractéristiques acoustiques (coefficients MFCC) extraits du signal de parole dans le domaine temporel. Le signal de parole est transformé en une série de vecteurs caractéristiques (vecteurs caractéristiques) avec des coefficients MFCC (Mel-Frequency Cepstral Coefficients). L'extraction des paramètres est réalisée avec l'outil wave2feat sphinx. La procédure pour créer le modèle acoustique consiste à regrouper un ensemble de données d'entrées et à les traiter avec l'outil

SPHINXTRAIN. Dans le contexte des systèmes de reconnaissance markovien, le modèle acoustique est généralement un HMM, typiquement un HMM gauche-droite à trois états appelé Bakis (voir figure 9), dans un état associé à un phonème. Nous avons conservé les paramètres par défaut du système CMU Sphinx comme système de base de notre travail.

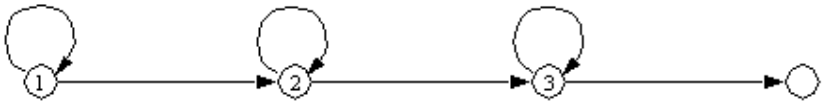

Figure 9 : HMM Bakis gauche-droite à trois états

5.5. Résultats

Dans nos expériences, nous avons utilisé deux ensembles de données : un pour l'apprentissage du système et l'autre pour les tests. L'indépendance du locuteur est accomplie dans nos expériences. Ceci est important car les systèmes de reconnaissance de la parole doivent répondre aux différences entre les locuteurs. Il est certain que tous les utilisateurs probables ne peuvent pas être utilisés dans l'apprentissage du système. Le système doit pouvoir s'adapter aux utilisateurs qui ne sont pas utilisés pour former le système. Dans ce travail, nous avons utilisé des données importantes pour développer le système qui le rend plus indépendant du locuteur (voir tableau 3).

Données	**Apprentissage du système**		**Test du système**		**Total**
	H	**F**	**H**	**F**	
Nombre de locuteurs	20	20	5	5	50
Nombre de locuteurs avec 5 répétitions.	40 * 5 = 200		10 * 5 =50		50 * 5 = 250
Total de 150 mots	150 * 200 = 30000		150 * 50 =7500		37 500

Table 3 : Total de fichiers utilisés pour l'apprentissage et test du système

Une base de données de 150 mots est sélectionnée du corpus amazigh développé dans le cadre de ce travail comme premier test de notre corpus et paramétrage du système de référence. Au total, 50 locuteurs ont prononcé les 150 mots amazighs avec cinq répétitions. Les 40 premiers locuteurs ont été utilisés pour l'apprentissage du système, les 10 autres étant réservés aux tests. 37 500 est le total de fichiers utilisé. Sur la base des valeurs par défaut de CMU Sphinx, les taux de reconnaissance obtenus pour la base de données est de 70%. A partir des résultats obtenus, on voit clairement que les résultats obtenus en utilisant des valeurs par défaut sont satisfaisants.

6. Conclusion

Ce travail est une contribution au traitement automatique de la parole amazighe, dans lequel nous avons étudié et réalisé un système de reconnaissance automatique de la parole en utilisant un environnement basé entièrement sur la langue amazighe. Dans ce cadre, nous avons construit une base de la parole amazighe composée de plus de 573 mots isolés prononcés par 50 locuteurs. 150 mots sélectionnés de cette base comme travail préliminaire ont été utilisés et évalués en développant un système de reconnaissance de la parole basé sur le toolkit CMU Sphinx. Le système de base développé a montré de bonnes performances, il utilise les modèles de Markov cachés comme classifieur. Il s'agit d'un travail préliminaire qui nous permettra d'accomplir notre objectif de base qui est un système de reconnaissance de parole indépendant du locuteur. Différentes valeurs peuvent être améliorées pour trouver la meilleure combinaison qui assure les meilleures performances pour le langage amazighe.

Références

Abenaou, A., Ataa Allah, F., Nsiri, B. (2014). Vers un système de reconnaissance automatique de la parole amazighe basé sur les transformations orthogonales paramétrables. Asinag n°9. pp. 133-145.

Abushariah, M. A. M., Ainon, R. N., Zainuddin, R., Khalifa, O. O., Elshafei, M. (2010). Phonetically rich and balanced Arabic speech corpus: an overview. IEEE Proceedings of the International Conference on Computer and Communication Engineering (ICCCE'10), Kuala Lumpur, Malaysia.

Al-Sulaiti, L., Atwell, E. (2006). The design of a corpus of contemporary Arabic. International Journal of Corpus Linguistics, John Benjamins Publishing Company. pp.136.

Ananthi, S., Dhanalakshmi, P. (2013). Speech Recognition System and Isolated word Recognition based on Hidden Markov Model (HMM) for Hearing Impaired, International Journal of Computer Applications. 73(20):0975-8887.

Boulaknadel, S., Ataa Allah, F. (2013). Building a Standard Amazigh Corpus. In: Kudělka, M., Pokorný, J., Snášel, V., Abraham, A. (eds) Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011), Prague, Czech Republic, August, 2011. Advances in Intelligent Systems and Computing. Vol. 179. Springer, Berlin, Heidelberg.

El ghazi, A., Daoui, C., Idrissi, N., Fakir, M.. Bouikhalene B. (2011). Système de reconnaissance automatique de la parole Amazigh à base de la transcription en alphabet Tifinagh, Revue Méditerranéenne des Télécommunications. Vol. 1, n°2.

El ghazi, A., Daoui, C., Idrissi, N. (2014). Automatic Speech Recognition for Tamazight Enchained Digits. world Journal Control Science and Engineering. 2(1):1-5.

Ghai, w., Singh, N. (2012). Analysis of Automatic Speech Recognition Systems for Indo-Aryan Languages: Punjabi A Case Study: International Journal of Soft Computing and Engineering (IJSCE). Vol. 2, Issue-1. ISSN: 2231-2307.

Hoge, H., Tropf, H.S., w inski, R., Van den Heuvel, H., Haeb-Umbach, R., Choukri, K. (1997), European speech database for telephone applications. in Proceedings of the IEEE ICASSP, European Language Resources Association (ELRA). http://www.elra.info/, Vol.3, pp.1771-1774.

Kayte, C. N., Pawar V. P., Sonawane, C. D. (2012). Human, computer interaction using isolated-words speech recognition system. Indian Streams Research Journal.

Kimutai, K., Milgo, E., Gichoya, D. (2013). Isolated Swahili w ords Recognition using Sphinx4, International Journal of Emerging Science and Engineering (IJESE). Vol. 2, Issue-2. ISSN: 2319–6378.

Kumar, K., Jain, A., Aggarwal, R.K. (2012), Hindi speech recognition system for connected words using HTK, Int. J. Computational Systems Engineering. 1(1):25-32.

Pokhariya, J. S., Mathur, S. (2014). Sanskrit Speech Recognition using Hidden Markov Model Toolkit. International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 10. ISSN: 2278-0181 IJERTV3IS100141.

Rabiner, L.R. (1989). A tutorial on hidden Markov models and select applications in speech recognition. Proceeding of IEEE, 77(2):257-286.

Robinson, T., Fransen, J., Pye, D., Foote, J., Renals, S., (1995), wSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition, in International Conference on Acoustics, Speech, and Signal Processing, ICASSP-95., 1(9-12):81-84.doi: 10.1109/ICASSP.1995.479278.

Sameti, H., Veisi, H., Bahrani, M., Babaali, B., Hosseinzadeh, Kh. (2011). A large vocabulary continuous speech recognition system for Persian language, EURASIP Journal on Audio Speech, and Music Processing 2011. http://asmp.eurasipjournals.com/content/2011/1/6.

Satori, H., El Haoussi, F. (2014). Investigation Amazigh speech recognition using CMU tools. Int J Speech Technol. https://doi.org/10.1007/s10772-014-9223-y.

Telmem, M., Ghanou, Y., (2018). Estimation of the Optimal HMM Parameters for Amazigh Speech Recognition System Using CMU-Sphinx, Proceedings of the first international conference on intelligent computing in data sciences, https://doi.org/10.1016/j.procs.2018.01.102.

Vrinda, M., Shekhar, C. (2013). Speech Recognition System For English Language, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 1.

ⵙⵓⵛⵑ++ⵓ ⵒⵏⵎⵙⵓ ⵕⵏ +ⵙⵍⵑⵓⵙⵍⵙⵏ ⵙ++ⵕⵍⵓⵓⵛⵍⵕⵕⵏ ⵠ ⵒⵓⵕ+ⵙ ⵏ +ⵍⵕⵕⵕ +ⵙⵓ +ⵕⵛ+ ⵏ +ⵙⵏⵕⵍ+ +ⵕⵠⵕⵕⵕⵎⵕⵏ+ ⵅⵃ +ⵛⵕⵅⵙⵎ+ ⵏ +ⵙⵏⵏⵒⵎⵏⵙ+ ⵏ ⵒⵏⵎⵛⵙⵓ ⵏ ⵒⵛⵎⵕⵍⵕⵛ.

ⵕⵓ ⵓⵕⵍⵕⵎⵏ+ +ⵍⵑⵓⵙⵍⵙⵏ ⵕⵏ ⵅⵃ ⵏⵠⵕⵛ ⵏ ⵙⵠⵓⵕⵏ : ⵕⵎⵛⵛⵒⵏ ⵓ +ⵙⵏⵏⵒⵎⵏⵙ+, ⵙⵓⵒⵠⵕⵛ ⵙⵛⵒⵒⵒⵒⵏⵏ ⵏ ⵙⵒⵒⵒⵒⵏ ⵏ +ⵒ+ⵎⵕⵎ+, ⵕⵓⵛⵏⵍ ⵏ ⵍⵕⵍⵕⵎ, ⵏ +ⵒⵏⵅⵕ +ⵕⵓⵙⵏⵏⵕⵕ ⵏ ⵙⵓⵏⵏⵠⵙⵎ.

ⵓⵒⵛⵏⵏ+ +ⵍⵑⵓⵙⵍⵙⵏ ⵕⵏ ⵕⵏⵓⵕⵍ ⵏ ⵙⵛⵒⵅⵕⵎ ⵏ ⵙⵛⵓⵅⵒ+ⵏ ⵏ ⵙⵅⵅⵒⵎⵏⵏ ⵙⵏⵕⵛⵒⵓⵏ ⵏ ⵙⵠⵠⵓⵏⵏ ⵠ ⵎⵙⵠⵓ ⵏ +ⵙⵏⵏⵒⵎⵏⵙ+ ⵏ ⵒⵏⵎⵛⵙⵓ ⵏ ⵒⵛⵎⵕⵍⵕⵛ ⵒⵒⵒⵒⵕⵕ ⵅⵃ +ⵒ+ⵎⵕⵎ+ⵎⵙ +ⵙⵠⵕⵛⵎⵙ, ⵎⵒⵛⵛⵏ +ⵒ+ⵎⵕ+ +ⵕⵛⵕⵅⵙⵎ+.

* * *

يتضمن هذا الكتاب، الأعمال التي قدمت في إطار الدورة الثامنة للندوة الدولية حول الأمازيغية وتكنولوجيا المعلومات والاتصال. وتتناول أربع مجالات رئيسة، وهي : التعلم بوساطة التكنولوجيا، والموارد والأدوات اللغوية، ومعالجة الكلام، والتعرف الضوئي على حروف تيفيناغ.

وتعرض هذه الأعمال مساهمات الباحثين والمهنيين، الوطنيين والدوليين، العاملين في مجال تكنولوجيا المعلومات والاتصال المطبقة على اللغات الطبيعية، وخاصة اللغة الأمازيغية.

* * *

Cet ouvrage contient les travaux de la 8ème édition de la Conférence Internationale sur les Technologies d'Information et de Communication pour l'Amazighe (TICAM). Ils abordent quatre grands axes, à savoir : l'apprentissage médiatisé par la technologie, les ressources et outils linguistiques, le traitement de la parole et la reconnaissance optique des caractères. Ces travaux pour état des contributions des scientifiques, chercheurs et professionnels nationaux et internationaux qui œuvrent dans le domaine des Technologies d'Information et de Communication appliquées aux langues naturelles, notamment la langue amazighe.

48 DH